非負値行列因子分解アルゴリズムに基づくメッセージ特徴の選択手法に関する研究

非負値行列因子分解アルゴリズムに基づく メッセージ特徴の選択手法に関する研究

筑波大学
図書館情報メディア研究科

2019年4月

輪島 幸治

非負値行列因子分解アルゴリズムに基づく メッセージ特徴の選択手法に関する研究

輪島 幸治

携帯型端末の小型化と高性能化の進歩で，発展してきた情報化社会は，伝播メディアやコミュニケーション方法，必要とされるリテラシー能力にも影響を与えてきた．特に，日本では 1985 年に登場したショルダーフォンに端を発したモバイル通信機器は，2013 年にはスマートフォンへと変遷し，電話だけでなくメールやチャットなどのテキストコミュニケーションを身近なものとした．また，インターネットと情報発信プラットフォームの登場で，これまでマスメディアが主役を担っていた情報発信は，個人やコミュニティへと開放され，一個人でも不特定多数に情報を発信できる環境が整ってきた．このような個人による情報発信に基づいて形成されるメディアは，ソーシャルメディアと呼ばれ，21 世紀を特徴付けるコミュニケーションメディアとなっている．

個人が情報発信の主体となるソーシャルメディアにおいては，情報は情報の発生源から直接大量かつ一方向で発信される傾向が強い．このため，従来のマスメディアが編集というプロセスを経て情報を選別し，発信してきたモデルとは大きく異なると言える．情報過多な時代の伝播メディアでは，受信者にとって，2 つの事象が発生する．第一の事象は，情報取捨選択の機会の増加である．情報を配信する伝播メディアが多いことから，不要なメッセージも増加する．第二の事象は，受信者の趣味や興味の多様化である．多様化が進んだ場合，パーソナライゼーションやレコメンドなどが行われる．この 2 つの事象から導き出される問題は，個々の受信者の都合に合わせた情報発信は一層難しくなってきていると言うことである．受信者の環境や受信者が多様化していることから，反応を予見することは困難である．受信者が不要と判断した伝播メディアからの情報は遮断され，その後は受信者への到達の機会を失う．結果，情報が伝達されるかは紙一重の状況となる．また，受信者の中には過剰な反応など，炎上のきっかけを作る受信者も少なからず存在する．したがって，発信者は受信者に有効な優れたメッセージ作成を行う必要があると言える．提案手法では，課題解決に有効な優れたメッセージを明らかにすることを目的としている．本研究では，課題解決に有効な複数の特徴量からなる高次な特徴をメッセージ特徴と呼ぶ．具体的には，2,000 次元を越える特徴量に対し，非負値行列因子分解 (NMF) による特徴量変換を適用し，変換特徴量であるメッセージ特徴を得る．そして，提案手法でメッセージ特徴を評価し，有効なメッセージ特徴を選択する．選択したメッセージ特徴における特徴量の寄与率に基づいて，課題解決に有効な少数の特徴量の集合を抽出する．優れたメッセージ特徴が明らかになることで，受信者に有効なメッセージに改善することが期待できる．

ここで，伝播メディアや課題に応じてメッセージは大きく異なる．したがって，複数の改善アプローチが必要である．本研究では，2 種類の改善アプローチを行う．まず，伝播メディアにおいては，メディアの種類が非常に大きな役割を持つ．したがって，メディアの種類に基づいた評価は有効である．また，テレビ放送における視聴率やラジオにおける聴取率など数値に着目した評価もこれまでと同様に重要である．このため，数値に基づいた評価は有効である．

本論文では，メディアの種類と数値に着目した 2 種類のメッセージ特徴の選択手法を提案している．第一の選択手法は，メディアの種類に基づいた評価を目的に，非負値行列因子分解アルゴリズムとグレゴリー・ベイトソンの情報の定義である「“違い” を生む “違い”」を組み合わせた手法である．これを差異に基づくメッセージ特徴の選択手法と称する．差異に基づくメッセージ特徴の選択手法は，閲覧数や返信率などの変数，また正解や不正解など明確な正解データが得られない場合においても，有効なメッセージ特徴の選択手法となっている．

第二の選択手法は，目的変数となる数値データが存在する場合を想定した，非負値行列因子分解アルゴリズムとサポートベクタ回帰モデルを組み合わせた手法である．これを回帰に基づくメッセージ特徴の選択手法と称する．

提案手法を実装評価し，分類タスクを用いた性能評価実験を用いて，提案手法の有効性を評価した．選択した特徴を用いた分類実験を行い，提案手法の有効性を検証している．本研究では，伝播メディアのメッセージ受信者が，主観的な印象を判断する要素に着目した．本研究における評価では話題，平易化，プライバシー，共感を取り上げている．結果，分類タスクで優れた分類精度を示し，有効性が明らかとなった．

本論文は，概要を述べた第 1 章を加えて全 6 章で構成されている．以下，各章の概要を述べる．第 2 章では，関連研究を示す．伝播メディアにおける課題である話題および共感，平易化，プライバシーに関する先行研究を概観し，既存研究に対する本研究の貢献を明らかにする．話題および共感に関する先行研究では，トレンドキーワードや情報カスケードなどの既存手法を述べ，これまでの研究トレンドを明らかにする．テキスト情報の平易化に関する先行研究では，コンテンツの難易を評価する研究とコンテンツを平易化する研究を概観している．プライバシーに関する先行研究では，プライバシーの多義性やプライバシー侵害などに関する既存研究を紹介し，これまでの動向を明らかにしている．

第 3 章では，本研究におけるメッセージ特徴とその有効性判別方法を示す．メッセージ特徴を得るための非負値行列因子分解アルゴリズムを示し，本研究で定義するメッセージ特徴を論じる．また，メッセージ特徴の有効性を判別するために使用する分類器，および，非線形回帰，グラフィカルモデルを示す．本研究では，Ada Boost，Random Forests，MLP(Multi-Layer Perceptron)，K-Nearest Neighbours(K-NN) の 4 種類の分類器を用いて提案手法の有効性を明らかにしている．分類精度の評価指標は，一般的に広く使われている適合率，再現率，F-measure を用いた．非線形回帰には，サポートベクタ回帰モデルを用いた．ここでの評価指標は，目的変数に対する予測誤差であり，MAE(MeanAbsoluteError) および RMSE(RootMeanSquaredError) の 2 つを用いた．グラフィカルモデルは，本研究において因果関係の推定に使用しており，本研究では，代表的なベイジアンネットワークを用いた．

第 4 章では，差異に基づくメッセージ特徴の選択手法を提案している．メッセージ特徴をグレゴリー・ベイトソンの情報の定義に基づき，異なるメディアの集合における係数値の差に着目し，メッセージ特徴を評価した．具体的な実装を述べた後に，性能評価実験として行った話題性および平易化における伝播メディアの課題を述べている．

“実験 1-1” では，質問記事の話題性に着目し，質問記事における閲覧数に基づいて評価した．閲覧数に基づいて評価することで，質問記事のコンテンツで，閲覧数の異常や変化の兆し，投稿後に話題になる質問記事を検知できるとした．評価実験の結果，クラス・メディアのオンラインコミュニティの分類タスクで優れた分類精度が得られた．また，非線形回帰を用いた回帰タスクにおいても，クラス・メディアのオンラインコミュニティで明瞭な予測精度の向上が得られた．分類タスクおよび回帰タスクで優れた結果が得られたことから，提案手法の有効性が確認できたと言える．

“実験 1-2” では，特別なスキルを求めない幅広い読者層に向けて発行した普及啓発書に基づいて評価した．普及啓発書に基づいて評価することで優れたメッセージ特徴が明らかとなり，特別なスキルを有しない幅広い読者層に向けてより良い情報発信が期待できるとした．評価実験の結果，優れた分類精度が得られた．また，因果関係分析においても，有効性が示された．分類タスクおよび因果関係推定で優れた結果が得られたことから，提案手法の有効性が確認できたと言える．

第 5 章では，回帰に基づくメッセージ特徴の選択手法を提案した．非線形回帰が行えるサポートベクタ回帰モデル用いて，目的変数を基づいてメッセージ特徴を評価している．回帰に基づくメッセージ特徴の選択手法では，性能評価実験として，プライバシーおよび共感における伝播メディアの課題を評価した．

“実験 2-1” では，プライバシー侵害に影響がある SNS 投稿記事をプライバシー侵害のアンケート結果に基づいて評価した．アンケート結果である数値に基づくことで共有されたコンテンツが，他者に不愉快な感情，あるいはプライバシー侵害に相当するコンテンツであるか否かを，コンテンツから判断できるかを明らかにすることを評価実験の目的とした．評価実験の結果，分類タスクにおいて優れた分類精度が得られた．また，因果関係推定においても複数基底の評価で有効な結果が得られた．分類タスクおよび因果関係推定で優れた結果が得られたことから，提案手法の有効性が確認できたと言える．

“実験 2-2” では，利用者が有益と判断するコンテンツを質問記事の返信数に基づいて評価した．返信数を用いることで，オンラインコミュニティにおいて影響が大きいコンテンツを判断し，情報推薦や情報の管理などで，利用者がより良く利用できるとした．評価実験の結果，分類タスクにおいて優れた分類精度が得られた．また，複数の基底評価を行った場合においても，提案手法は優れた結果が得られた．加えて，非線形回帰を用いた回帰タスクにおいても，有効な結果が得られた．分類タスクおよび回帰タスクで優れた結果が得られたことから，提案手法の有効性が確認できたと言える．

第 6 章で，本研究における結論を示す．結論では，本研究における総括を述べ，提案手法における今後の課題と，成果を踏まえた望ましい発展の方向について，私見を交えた展望を示している．

A Study on Characteristic Selection in the Messages based on Non-negative Matrix Factorization

Koji Wajima

The media of dissemination (Verbreitungsmedien) have rapidly changed owing to technological progress, especially in information and communication technologies. Reflecting the changes in the conditions of technological progress, communication methods and abilities have also changed. Consequently, portable terminal equipment has advanced from shoulder phones to smartphones. Currently, online communication that is independent of time and place is increasing. Simultaneously, social concerns associated with online communication have been increasing over the last several years. Information regarding online communication significantly impacts purchasing behavior in consumer generated media. For example, inappropriate online behavior leads to "Enjyo" and "Framing," causing issues in online media. Therefore, message analysis in the media of dissemination is necessary.

In the media of dissemination, two phenomena have been observed in recent years. The first phenomenon is the increase in opportunities associated with selecting the information of receiver and the increase in the various distribution modes available. For a receiver, the information of the unnecessary spread media is not read (ignore). Hence, the transmission efficiency of information decreases. The second phenomenon is the diversification of the receivers' hobbies. Customization in terms of the information received is realized through this diversification. Therefore, for a sender, assuming that the receiver has received and read the transmitted information is difficult.

Consequently, media of dissemination need to work improvement of the message. In this study, message characteristics were extracted to improve message quality in the media of dissemination. If the extracted characteristics are superior, then important characteristics are revealed. Therefore, message characteristics contribute to message distinction. The improvement plan can be separated with respect to two aspects: the media and the objective variable. Messages transmitted via different types of media differ; thus, the type of media plays a significant role. Therefore, message evaluation is based on the different types of media. In addition, numerical evaluations are important. Different media involve various metrics, such as audience ratings. Therefore, message evaluation is also based on numerical values. In this study, I evaluated message characteristic selection based on Berlo's sender-message-channel-receiver (SMCR) model, wherein the evaluation targets were text messages. Berlo's SMCR model is a standard communication model that can use online text messages for applications such as ethology communication, electrical signal communication, and engineering communication.

The proposed method comprises extracted, converted, and base-evaluated feature quantities. First, characteristics were extracted from the message. Note that the evaluation results may change when a specific characteristic is included in the given message. For example, there may be a paradox in a sentence or a phrase. In this study, I used multivariate analysis to decompose a message into additive components. In the case of a component, the paradox of the given sentence is considered. Therefore, the occurrence of problems decreases. The proposed method uses non-negative matrix factorization (NMF). In existing research, NMF has demonstrated superiority over other multivariate analysis techniques. In addition, algorithmic expansion and improvement have been realized by the research community. As a result, NMF is the most suitable algorithm for our purpose. The NMF base result is the result of the group of co-occurrence ingredients of the quantity of the characteristics. Here, the NMF base was evaluated using methods based on Bateson's definition of information and nonlinear regression. Note that the text characteristics have a contribution ratio based on each NMF base; therefore, characteristics with high contribution are based on the characteristics of the co-occurrence of the NMF base. The characteristics of the co-occurrence correspond to message characteristics, and message characteristics that are good for problem solving are required. Therefore, the result is selecting the appropriate characteristic for the problem of the media.

The proposed method was experimentally evaluated using online text messages. First, I extracted the feature from the media. Then, I employed feature conversion via NMF. Note that all extracted characteristics were used in the NMF. Finally, the NMF base was evaluated based on the research subject. In this study, the evaluation experiment used Japanese text characteristics (surface layer information, topic, word type, basic vocabulary, semantic attributes, verbal expression, sentence end expressions, part of speech type, unique expressions, and evaluation expressions). In this experiment, a substantial amount, i.e., 31 types of characteristics and 2,071 dimensions, of characteristic data from previous studies was considered. In addition, another experiment was conducted using 33 types of characteristics and 2,073 dimensions. The research subject was Japanese text characteristics (including topic, simplified corpora, privacy, and empathy). Several research subjects were evaluated using the proposed method. High classification performance was observed using the extracted message characteristics. Furthermore, it was demonstrated via an evaluation experiment that the proposed method can fit the problems of various media. The results of this experiment will change the characteristics of the online text message.

The remainder of the manuscript is organized as follows. Section 1 outlines the research and the research subject. Section 2 describes related work. Section 3 describes the proposed method for classifying online text messages based on non-NMF. In addition to the classifier, nonlinear regression, and Bayesian network.

Sections 4 and 5 describe the proposed method and an evaluation experiment, respectively. Section 4 also describes "characteristic selection in the text messages based on differences." The evaluation experiment of the proposed method comprises the topic of the online community and simplified corpora of text information. The topic evaluation objects used herein were online communities of two types of media. The proposed method demonstrates superior results as compared with document classification and nonlinear regression methods. The evaluation target for simplified corpora was the Japanese government's "Annual Report on the Environment." The proposed method also demonstrates superior results as compared with yearly comparison and Bayesian network methods. Section 5 discusses "characteristic selection in text messages based on nonlinear regression." The evaluation experiment of the proposed method consider the privacy of SNS articles and the empathy of media. The evaluation object for privacy was the SNS article. The proposed method demonstrated superior results compared to document classification and Bayesian network methods. The evaluation objects related to empathy were online communities from two types of media and four different types of media. The proposed method also demonstrated superior results as compared with the document classification method, multi-base evaluation, and evaluation of different media.

Section 6 presents conclusions and suggestions for future work. I expect that the results of this study will improve the characteristics of the message. In addition, message characteristics would strengthen a powerful effect or the influence of the media. In the future, other message types will be considered relative to other research areas. In addition, owing to this line of research, similarity problems can be solved in principle if they are converted to a vector space model. Further research on message characteristics would clarify need message. It is expected that unnecessary media and messages will be removed. I expect it to contribute to mutual future development. In the future, additional experiments will be conducted using the proposed method with other message types. In addition, the extraction of new features will be required for other message types. It is my hope that the results of this study will contribute to the development of a better information-oriented society.

目 次

図 目 次

表 目 次

第1章　序論

1.1　背景

近年，情報化社会の発展でオンライン上のコミュニケーションが増加している．20 世紀は機械を中心に科学技術が飛躍的に発展されてきた．21 世紀では，科学技術に基づく，人間主体の発展となるとされている [1]．人間主体の発展において，コミュニケーションで，重要視されるのは，情報伝達対象に対して，有益な情報伝達を行うことである．

伝統的なメディア呼ばれる媒体に，マスコミ 4 媒体がある．マスコミ 4 媒体は新聞，雑誌，ラジオ，テレビの 4 つの媒体である．各地に拠点を持つことは，ニュース・ネットワーク [2] [1] やプレス・リレーションズ [2] [2]によって，各地域で生活する人々の特性を含んだ実体や 20 年〜30 年の継続した長期の取材が行える [3]．新聞，雑誌，テレビなどは，社会制度のように社会を形成し，維持する仕組みに基づいている．したがって，スタティックな性質を持つ．

伝統的なメディアは世帯から個人へ，大衆から個へという潮流があるが [4]，一方で，“分衆・小衆論”，“テレビ離れ論” など呼ばれ，同様の議論は以前より行われてきた [5]．マスコミ 4 媒体に掲載される広告をマス広告と呼ぶ [2]．日本における総広告費は，2007 年度において 7 兆円以上 [2] であり，40% から 50% が，マスコミ 4 媒体である伝統的なメディアにおけるマス広告費である．テレビを始めとした伝統的なメディアは各国や全国各地に通信社や民放各社系列局を持つ．2007 年度において，民放各社の系列を含め 127 社，民放ラジオ局においては 101 社あり，系列社数や広告費の規模から，伝統的なメディアの影響が大きいことが明らかである [2]．

影響が大きい伝統的なメディアは拠点を持つことで，共通性，共同性，連帯性といった地域性を超えた共同性を持つ．地域性を超えた共同性に，メディアが作る「現実 - 像 (image of reality)」であるベネディクト・アンダーソン[3]による “想像の共同体” という概念がある [6]．想像の共同体を持つことで，伝統的なメディアは課題設定機能という重要な課題を設定・議論し，民意を構築するといった役割を担ってきた [7]．伝統的なメディアで議論されている課題には “世代間公正問題”，“公的年金と国債”，“苦情・退出・愛顧”，“若者への公正さ” などがある [8]．少子高齢化やライフサイクル仮説に基づく経済行動，世代間の課題など公共政策に関する議論は，伝統的なメディアによる “課題設定機能” の役割である [9]．

[1] マスコミ・報道機関との関係を密接にし，相互理解を進め好意的な報道を期待する活動，記者発表会，懇談会，取材対応など．

[2] 東京キー局を核とし，ニュースの素材提供，報道番組の共同制作，取材諸経費の分担などを目的に形成されたネットワーク．

[3] アメリカの政治学者・歴史家 (1936-2015)，近代国家はマスメディアが支える想像の共同体であると説いた．

伝統的なメディアは不特定多数の大衆に対して，大量の情報を伝達する．テレビなどの伝統的なメディアの放送分野では中正[4] であることが必要とされている．昨今において，世の中の動きについて信頼できる情報を得るメディアは 2000 年から 2015 年の推移においてもテレビが 1 位である [4]．ゆえに，伝統的なメディアの規模や影響力は非常に大きい．

しかし一方で，趣味[5]娯楽[6]の情報源として利用するメディアは，2000 年から 2015 年の推移で 1 位が “テレビ” から “インターネット” に変化した．これは情報化社会が進んだ結果であると言える．

図 1.1: メディアの変化と受容

インターネットの登場により，メッセージを送信するメディアは増加し，受信者の趣味や興味も多様化した．メディアが変化することは，情報伝達の速度や分量のような量的側面だけでなく，伝達内容といった質的側面，人間の能力面にも変革をもたらす．情報の伝達の速度が上がり，大量の情報から重要な情報を聞き分ける能力は向上する．一方で，情報アクセス機器に頼る結果，漢字が書けなくなるなど，能力の低下なども発生する可能性がある．

加えて，趣味娯楽の情報源であるインターネットは多くの世代で利用されている．したがって，誰もが閲覧できる環境であり，不特定多数のユーザが利用している．このため，インターネットのメディアは，ダイナミックな性質を持つ．インターネットにおいては，新しい形式に基づいた伝統的なメディアとは異なるアマチュアの想像の共同体も形成されている [10]．アマチュアの想像の共同体は，Global，誰でものようなオープンな性質を持つ．したがって，適切な対応を取り続けなければ炎上するなどの社会的な影響も大きい．

[4]特定の考えや立場に偏ることなく公正であること．
[5]職業や専門としてではなく，個人が楽しみで愛好していること．
[6]仕事・勉強を離れた余暇などに心を楽しませたり，慰めたりしてくれるもの．

1.2 問題設定

さて，メディアは，影響力の大きさや経済規模から，批判的な言説が多くなされるが，一方では，秩序を与える調節機構である．理論に基づいた場合，ニクラス・ルーマン[7]の理論社会学においては，到達の確実性としての伝播メディア[8]，適切な問題領域にコミュニケーションを導く成果の確実性としての成果メディアに分類されている [11].

情報化社会の発展で，科学技術に基づいた場所に依存しないコミュニケーションの増加した．20 世紀は機械を中心に科学技術が発展し，情報アクセス機器の普及や伝播メディアの増加，そして，インターネットが登場した．インターネットで，チャット・メール，ソーシャルメディアが登場し，受信者にとって，情報過多な状況が発生する．21 世紀では，科学技術に基づく人間主体の発展である [1]．したがって，受信者に効果的な情報伝達を行わなければならない．情報過多な時代の伝播メディアでは，受信者にとって，2 つの事象が発生する．1 つ目は，情報の取捨選択の機会の増加である．伝播メディアが多いことから，受信者にとって，不要な伝播メディアの情報は閲覧されない．結果，伝播メディアに対する受信者の反応は 2 極化する．したがって，発信者にとっては，受信者の想定は困難な状況となる．

2 つ目は，受信者の趣味や興味の多様化である．趣味や興味が多様化することから，過剰な反応や炎上が発生する．結果，伝播メディア媒体の多様化が進む．このため，情報が伝達されるかは紙一重の状況となる．この 2 つの事象から導き出される問題は，受信者の都合に合わせた情報発信は困難になっているということである．ゆえに，マス・コミュニケーション[9]のように，多くの人々に向けて行われるコミュニケーションは，情報発信における改善が必要とされている．

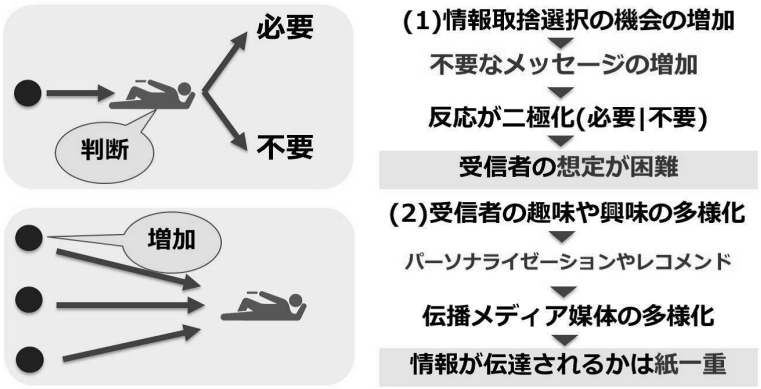

図 1.2: 情報過多な時代の伝播メディア

[7] ドイツの社会学者 (1927-1998)，行為よりもコミュニケーションを単位として社会システムを捉え，権力・信頼など幅広いテーマを論じた．

[8] 物理的な伝播メディアは，文字，画，電磁波など．社会制度を含む場合の伝播メディアは，郵便，テレビ放送など．伝播 (でんぱ) は，波動が広がっていくことを意味する．

[9] 新聞，雑誌，図書，映画，テレビ，ラジオなどのマス・メディアを通して，多くの人々に向けて行われるコミュニケーション．

ここで，情報発信における改善では，伝播メディアにおけるコミュニケーションに基づいて評価しなければならない．コミュニケーションが行われる物理的な範囲 (空間・時間・距離など) は受け手の基本属性 (世代・社会的関心・消費行動など) は異なる．現実社会においては，多種多様な伝播メディアが存在する．本研究では，古典的なコミュニケーションモデルの一つである通信系のモデルを拡張したデイビット・K・バーロ[10] の SMCR モデルに着目した [12]．SMCR モデルは，情報源からチャネルを介したメッセージで，受信者に伝達するコミュニケーションを対象としている．また，電気機械的信号や音声など，通信工学も包含した通信系のモデルであり，伝播メディアに限らず，多様なコミュニケーション方法を論じることが可能である．SMCR モデルの要素は,「情報源 (Source)」「メッセージ (Message)」「チャネル (Channel)」「受信者 (Receiver)」で構成されている．伝播メディアにおける SMCR モデルの適用例を図 1.3 に示す．受信者の都合に合わせた情報発信は困難であることから，今後は，経路となるチャネル (Channel) や媒体に対する受信者 (Receiver) における評価よりも，相対的に情報源 (Source) が作成するメッセージ (Message) が重要となってきていると言える．

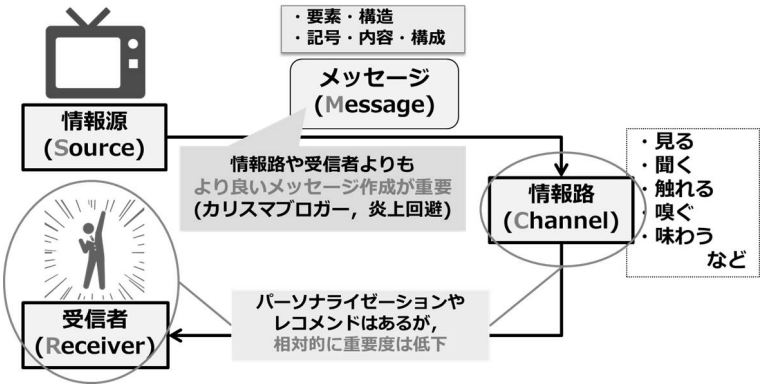

図 1.3: D.K. バーロの SMCR モデル [12]

本研究では，優れたメッセージを作成するために 2 種類の改善アプローチを行う．まず，伝統的な伝播メディアであるマス・メディアをオーディエンスの態様から分類した概念に，ゼネラル・メディア[11] とクラス・メディア[12] がある [2]．メディアの種類に応じて伝達されるメッセージは異なる．伝播メディアにおいては，メディアの種類が非常に大きな役割を持つ．メディアの種類に対して，テレビ放送における視聴率[13] やラジオにおける聴取率[14] など，数値に着目した評価もこれまでと同様に重要である [2]．数値に着目した評価を行うことで，情報の伝達状況や反応が評価できる．本研究では，伝播メディアにおけるメディアの種類と，数値の 2 種類の課題を解決する方法として，メッセージ特徴を用いる．

[10] アメリカの通信理論学者 (1929-1996)，シャノン理論の通信系のモデルを拡張した SMCR モデルを作成．
[11] 一般日刊紙やテレビなど，教育程度，職業，ライフスタイルなど，あらゆる社会全般の人々を普遍的にオーディエンスにしている媒体
[12] 雑誌・ラジオ，業界紙や専門雑誌，ダイレクト・メールなど特定の集団や限定された対象を相手にする媒体
[13] ある放送番組が，一定の地域で，どれくらい世帯にどの程度見られていたかを示す割合．
[14] ラジオ放送がどのくらいの聴取者に聞かれているかを示す割合．

1.3 本研究のアプローチ

本研究における提案を示す．本研究では，優れたメッセージを作成するために，高次な特徴量の集合であるメッセージ特徴を抽出する．優れたメッセージ特徴が明らかになることで伝播メディアの優れたメッセージ作成が期待できる．

これまでの研究において，伝播メディアにおけるメッセージでは，特定の特徴があった場合においても，逆接関係を導く表現が存在した場合，以降の関係において，反転する場合があることなどが，既存研究において指摘されている [13]．本研究では，コミュニケーションにおけるメッセージから，課題解決に有効な複数の特徴量からなる高次[15]な特徴を抽出することで，目的達成を試みた．本研究では，高次な特徴をメッセージ特徴と呼ぶ．第 3 章にて詳述するが，本研究における提案は，特徴量変換アルゴリズムである非負値行列因子分解 (NMF) を用いて特徴量の観測行列を構成成分に分解し，構成成分であるメッセージ特徴を評価する．そして，選択したメッセージ特徴における寄与率上位の特徴量を用いて，性能評価実験を行う．本研究では，大きく分けて 3 つの要件を実現する．

本研究における設定要件

要件 1. 受信者に有益と判断されるメッセージを判別できること

メディアの種類は非常に大きな役割を持つ．提案手法で得られるメッセージ特徴が，受信者に有益と判断されるメッセージを判別できることで，有益と判断されるメッセージに共通するメッセージ特徴が明らかになる．メッセージは特徴量の集合である．優れた特徴量の集合が明らかになることで，伝播メディアのメッセージ作成の改善が期待できる．提案手法および性能評価実験は，第 4 章にて詳述する．

要件 2. 目標変数に影響が大きいメッセージを判別できること

評価基準となる数値は重要である．提案手法で得られるメッセージ特徴が，目標変数に影響が大きいメッセージを判別できることで，受信者の反応に影響が大きいメッセージに共通したメッセージ特徴が明らかになる．メッセージは特徴量の集合である．優れた特徴量の集合が明らかになることで，伝播メディアのメッセージ作成の改善が期待できる．提案手法および性能評価実験は，第 5 章にて詳述する．

要件 3. メッセージ特徴が，伝播メディアの課題解決に有効であること

メディアの種類や媒体，メッセージを作成する情報源で，コミュニケーションにおけるメッセージは異なる．本研究では，提案手法の性能評価実験で，異なる 4 種類の伝播メディアの課題を評価する．4 種類の課題で優れた結果が得られることで，提案手法が幅広いコミュニケーションに対応した伝播メディアの課題への適応性を示すことが期待できる．第 4 章および第 5 章の 4 種類の性能評価実験で示す．

[15]数学で，3 次元より高い一般次元の多様体 (集合体) の総称．多様体は，幾何学的な類比を通じて，4 次元以上の空間を研究するために作られた概念．(x, y, z, t) は 4 次元多様体．幾何学は図形および図形の占める空間の性質について研究する数学の分野．類比は 2 つ以上の関係や機能が互いに類似していること，および未知のことを推し測ること．

1.4 伝播メディアにおける課題

1.3 節にて示したが，情報の取捨選択の機会の増加と，趣味や興味が多様化で，受信者の都合に合わせた情報発信は困難になっている．したがって，コミュニケーションにおいては，情報源 (Source) が作成するメッセージ (Message) が重要であると言える．多くの人々に向けて行われるコミュニケーションは，情報発信における改善が必要であることから，共通する要素に基づく，課題である必要がある．ここで，マス・コミュニケーションにおける 3 要素を示すと，事実の報道，解説・啓発，娯楽である [2]．この 3 要素は，マス・コミュニケーションの性質を持つ，すべての伝播メディアにおいて共通な要素であると言える．また，3 要素の要素に加えて，マス・コミュニケーションにおいては，放送受信者等の個人情報保護に関するガイドラインなどが設定されている．したがって，共通する要素に加えて，プライバシーに関する情報には配慮する必要がある．そこで，マス・コミュニケーションにおける 3 要素にプライバシーの要素を含めた 4 要素に基づいて課題解決を行う．

本研究では，受信者が主観的な印象を判断する基となる要素として，話題，平易化，プライバシー，共感を取り上げる．まず，伝播メディアにおけるメッセージでは，事実の報道を行う．メッセージは，多くの人に情報伝達することが目的である．伝播メディアにおいては，内容が刺激的であれば，情報伝達時に一定の影響力を持つ．したがって，伝播メディアにおける話題性は重要である．また，伝播メディアにおけるメッセージには，解説・啓発の要素が包含されている．伝播メディアにおいては，理解できない受信者が多い場合，影響は限定的である．受信者が理解できなければ，内容は伝達されない．このため，伝播メディアのメッセージの平易化は重要である．第 4 章にて後述するが，本研究においては，差異に基づくメッセージ特徴の特徴選択手法を用いて，話題性と平易化に関するメッセージ特徴を評価した．

次に，伝播メディアにおけるメッセージでは，プライバシーに関する情報には配慮する必要がある．受信側にプライバシー侵害などの不利益があった場合，不利益がある伝播メディアのメッセージは不要と判断する．したがって，伝播メディアのメッセージでは，プライバシー侵害となる情報は抑制することが重要である．そして，伝播メディアにおけるメッセージでは，娯楽の要素が包含されている．伝播メディアのメッセージを受信した場合でも，興味を持たない場合などは，受信者は反応しない．興味を持たない場合，伝播メディアあるいはメッセージを改善する必要があると言える．このため，伝播メディアにおける共感は重要である．第 5 章にて後述するが，本研究においては，回帰に基づくメッセージ特徴の特徴選択手法とその性能評価を用いて，伝播メディアにおけるプライバシーと共感に関するメッセージ特徴を評価した．

1.5 論文の構成

以下，本論文の構成を述べる．本論文は 6 章から構成されている．まず，第 2 章で関連研究を示し，本論文での評価実験の位置づけを明確にする．第 3 章では，本研究におけるメッセージ特徴とその有効性判別方法を示す．そして，第 4 章および第 5 章にて，メッセージの選択手法と伝播メディア課題への応用について論じる．その後，第 6 章で，本論文の総括を述べ，提案手法の課題と展望を示す．各章の概要を述べる．

第 2 章で関連研究では，本研究で伝播メディア課題として取り上げる，話題性および共感，テキスト平易化，プライバシーを論じる．そして，本研究において評価実験で用いる特徴量と既存研究における特徴量選択手法を示す．

第 3 章では，メッセージ特徴を得るための非負値行列因子分解アルゴリズム，本研究で定義するメッセージ特徴を示し，メッセージ特徴の有効性を判別するための分類アルゴリズム，非線形回帰，グラフィカルモデルについて論じる．

第 4 章では，提案手法であるメッセージ特徴の選択手法の一つである，差異に基づくメッセージ特徴の特徴選択手法とその性能評価として，特徴選択手法を示し，性能評価実験である”実験 1-1”および”実験 1-2”を述べる．“実験 1-1”では，オンラインコミュニティにおける質問記事の話題性を評価する．評価で，話題があるメッセージ集合に特有のメッセージ特徴を明らかにする “実験 1-2” では，年次報告書を用いて，テキスト情報における平易化を評価する．評価で，平易化なメッセージ集合に特有のメッセージ特徴を明らかにするそして，“実験 1-1”および “実験 1-2”から得られた結果を示す．

第 5 章では，提案手法であるメッセージ特徴の選択手法の一つである，回帰に基づくメッセージ特徴の特徴選択手法とその性能評価として，特徴選択手法を示し，性能評価実験である”実験 2-1”および”実験 2-2”を述べる．“実験 2-1”では，SNS におけるプライバシー侵害を評価する．評価で，プライバシー侵害のメッセージ集合に特有のメッセージ特徴を明らかにする “実験 2-2” では，オンラインコミュニティにおける共感を評価する．評価で，共感されたメッセージ集合に特有のメッセージ特徴を明らかにするそして，“実験 2-1”および “実験 2-2” から得られた結果を示す．

第 6 章では総括と今後の課題，展望を示す．総括では，提案方式を用いて得られた高次な特徴であるメッセージ特徴が，現実問題の解決に寄与することを示す．今後の課題では，提案手法の適用範囲を示し，課題となる部分を論じる．展望では，本研究の成果を基にした発展や応用を期待する分野に関して私見を述べる．

第2章　関連研究

本章では，話題および共感，平易化，プライバシーに関するこれまでの既存研究の概略を示す．そして，既存研究におけるメッセージにおける特徴量および特徴量の選択手法を示し，本研究における課題解決のアプローチを示す．

2.1　話題および共感

オンライン上のメディアを対象とした話題や動向の評価は，多くの研究が行われてきた [14][15]．流行や人気に影響を与える要素には，クチコミ[1] [16] や共感[2] [17]，また，技術のハイプ・サイクル (Gartner Hype Cycle) [3] [18] など技術トレンドを取り扱う研究から，Web 検索エンジンなどのクエリログを用いた周期性の発見 [19][20][21] に関する研究などもある．

話題性の評価のための手法には，トピック，トレンドキーワード (流行語)，ランキング，ソーシャルブックマーク，情報カスケードの分析などがある．トピックを用いた手法は，アルゴリズムでテキスト情報からトピックと呼ばれる単語集合を抽出し，時系列や種別に基づいて流行や人気を評価する [22]．トピックを用いた研究には，特定の話題が起因して，トピックが急上昇する現象であるバーストの研究 [23][24] や，話題の地域性を考慮することで精度が向上する報告 [25] などもある．トレンドキーワードとは，検索エンジンやソーシャルメディアにおいて，利用者から興味関心が高いキーワードである．検索エンジンの検索語であるクエリ[4] の頻度などで抽出が行われる．トレンドキーワードのウェブリソース間の振る舞いに関する研究や，コミュニティにおける発言割合の研究，ブログユーザ (ブロガー) への話題伝搬に関する研究などがある [26][27][28]．ソーシャルブックマークとは，編集したブックマークをインターネット上に公開できるウェブサイトである．ブックマークの周期性の分析や検索の時期，ソーシャルブックマークを用いた検索結果のランキング手法の研究がある [29][30]．

[1]友人・隣人などインフォーマルな情報源からの情報伝達である．購買決定においては，非営利的な人的情報源に分類されている．インターネットでは，各種商品やサービスなどに関する利用者 (消費者側) の評価・体験談の投稿など．

[2]他人の体験する感情を自分の体験のように感じること．

[3]Gartner - Research & Advisory Overview : https://www.gartner.com/en/research/methodologies

[4]データベースの検索で，指定された条件を満たす情報を取り出すために行われる処理の要求．

情報カスケード[5]とは，人々が何かに対する価値判断を行う際に，個々人が有する判断とは独立した状態で，集団全体が画一的な判断になだれ込んでしまう現象である．情報の拡散現象に対する予測や，社会的に影響力の判別の研究がある [31][32]．また，ユーザの影響力を評価する研究 [33] や，カスケードの将来の大きさを予測する研究 [34]，実際につぶやくユーザを予測する研究などもある [35]．

これらの研究に共通する要素は，話題を定義する特徴に基づいて，動向や周期性の評価を行っている点である．マイクロブログにおけるスパム検出 [36] や大規模な社会現象の分析 [37] などにも用いられている．加えて，オンラインコミュニティにおける話題は，言及されると，連動するように売上が増加する連動性もある [38]．評価基準においては，盛り上がりの早さや平均返信数 [39]，ユーザ行動やコミュニティの成長率 [40] などが用いられている．

ここで，話題性における動向や周期性を分析するためには，テキスト情報より，統計量や重要語などの要素を抽出し，出現頻度や推移に基づいて評価が行われている [41]．ゆえに，伝播メディアにおけるコミュニケーションにおいては，メッセージで受信者に必要，あるいは好意的な反応が得られるメッセージであることが重要である．このため，伝播メディアにおけるメッセージでは話題となる有益な特徴を明らかにすることが必要である．

2.2 平易化

平易化は，テキスト情報のコンテンツの難易を評価する研究である．テキスト情報の可読性を評価する研究と，テキスト情報を平易化する研究に大別される．

可読性とは，テキスト情報の読みやすさの度合いである．可読性には，コンピュータによる文書構造の認識しやすさである機械可読性と人間可読性があるが，ここでは人間可読性の既存研究を示す．可読性の研究は 1923 年より研究が行われている [42]．読みやすさの指標には，Flesch Reading Ease Score，Flesch-Kincaid Grade Level Score，The New Dale-Chall Readability Formula，Coh-Metrix などがある [43][44]．読みやすさの指標の特徴量には，文字の長さ，単語数，ひらがな比率などが特徴量に用いられている [45][46]．可読性に関する研究には，英語教材の開発の研究や，語彙のカバー率と理解度分析に関する研究がある [47][48]．平易化とは，難解な文書を平易化な文書に変換することである [49][50]．既存研究の特徴量には，文の長さ，語数，語の複雑さが使用されている [51][52][53][54][55]．

ここで，伝播メディアのコミュニケーションにおいては，受信者が理解できるメッセージである必要がある．したがって，可読性や平易化が重要である．このため，伝播メディアにおけるメッセージでは平易化された有益な特徴を明らかにすることが必要である．

[5]カスケードは，階段上に連続した滝の意である．

2.3 プライバシー

プライバシーの概念には，多義性と文脈依存性があることが知られており [56]，特定の単語が出現するか否かだけではプライバシーに言及しているかを論じることはできない．プライバシー保護の原則に，OECD プライバシー・ガイドラインに基づいて 1982 年に行政管理庁のプライバシー保護研究会が公表したプライバシー保護研究会報告書の 5 つの基本原則がある [57]．基本原則は，(1) 収集制限の原則，(2) 利用制限の原則，(3) 個人参加の原則，(4) 適正管理の原則，(5) 責任明確化の原則である．また，プライバシーの概念とされている権利や情報には，私生活をみだりに公開されない権利 [58] や肖像権 [59]，パーソナルデータなどがある [60]．基本原則に則ってプライバシー保護に関する様々な研究 [61] がなされてきた．

プライバシー保護に関する研究目的の一つにプライバシー侵害の発生防止がある．プライバシー侵害は Solove のプライバシー類型論に基づき，情報収集，情報処理，情報拡散，侵襲に大別することができる [62]．プライバシー保護に関する既存研究では，プライバシーに関連する投稿の検索 [63] など情報収集に関する研究，情報開示抵抗感・二次利用侵害懸念の調査分析 [64] など情報処理に関する研究．また，プライバシーの漏洩検知と公開範囲の設定 [65] など情報拡散に関する研究として広く知られている．特に情報拡散の研究では，個人のプライバシーに関する情報が流用・歪曲される場合が多いことから，漏洩検知と公開範囲の設定 [65] だけでなく，プライバシー侵害シーンの抽出 [66][67][68] など多岐にわたる．

また，SNS では情報の収集，処理，拡散が容易に行えることから，炎上やフレーミングと言う事象がしばしば発生している．ここで，炎上とは，コミュニケーション相手と異なる第三者が，逸脱や不適切と判断されるメッセージを問題視し，SNS 上で指摘することに端を発し，批判が集中する事象である [69]．一方，フレーミングとは，コンピュータ上のコミュニケーションで当事者のいずれかが敵対的で攻撃的な相互行為を取る事象である [70]．炎上やフレーミングの多くは SNS のコミュニケーションで過激な発言や行為を繰り返す事象であり，いじめ，嫉妬，批判などによって，冷静さを失った投稿者が行うことが多い．

インターネットでは不特定多数のユーザが利用しており，ソーシャルメディアでは，伝統的なメディアとは異なる想像の共同体が形成されている．特に，密なコミュニケーションが形成されている場では，クレームや批判的なコメントは急激に広がる．また，批判的なコメントによる炎上やフレーミングなどにより，他者のプライバシーを侵害，誹謗中傷に加担するなどプライバシーの問題も横行する [71]．意図したか否かに関わらず，プライバシーなど個人情報に関する内容については，侵害，誹謗中傷などで，社会的な損失も無視できない規模の問題へ発展する場合がある．メディアにおける課題はインターネットに限らず発生する．しかし，インターネットのように，不特定多数によって形成されるメディアの場合，構造や偏在化など受信者の集合の特性が起因し，発生する課題もある．

ここで，伝播メディアのコミュニケーションにおいては，受信者に与えるネガティブな要素に相当するメッセージを抑制する必要がある．したがって，伝播メディアにおけるメッセージではプライバシーに影響がある特徴を明らかにすることが必要である．

2.4 特徴量と特徴量選択に関する研究

2.4.1 テキスト情報の特徴量

本研究の評価実験で用いるテキスト情報の特徴量を “表層の特徴”，“話題 (トピック) の特徴”，“意味情報の特徴” に大別して簡単な説明を行う．

“表層の特徴” には，文の数，読点・句点の数，文の長さ，文字種類 [72]，読点間の距離 [73]，漢字包含率 [73] などがある．また，本研究の “表層の特徴” を用いた既存研究には媒体分析 [74] や書き手評価 [75]，文献の判定 [76] がある．“話題 (トピック) の特徴” には，アルゴリズムは，似た語彙の集合である話題を抽出する研究がある [22]．代表的な話題抽出のアルゴリズムには，LSI[77] や LDA[78] がある．本研究の “話題 (トピック) の特徴” を用いた既存研究には LSI は文書分類 [79]，LDA はユーザ推薦 [80] がある．

“意味情報の特徴” では，既存研究の辞書の研究を説明する．意味情報の抽出は，畳語 [81] など個別の特徴量抽出の研究が行われているが，テキスト情報の特徴量は多義性があり一意的でない．本研究では，理論社会学のゼマンティクの概念に基づき，既存研究で作成された辞書を用いる [82]．意味の特徴量は 8 種類に大別される．語種の特徴量は和語や漢語，外来語など語彙を分類した種類である．対象読者層の年代差評価 [83] の研究で用いられている．基本語の特徴量は使用率が大きく，使用範囲が広い語彙である [84]．ニュースの語彙分析 [85] で用いられている．意味属性の特徴量には，意味分類コードがある．意味分類コードは語を意味に基づいて分類したコードである [86]．文書の自動分類 [87] で用いられている．言語表現の特徴量は，機能語や複合辞などの機能表現がある [88]．価値判断の解析 [89] で用いられている．文末表現は，日本語の文章の固定化された文末表現 [90][91] である．ベストアンサーの推定 [92] や，文書の内容分析 [93] で用いられている．品詞は，日本語のテキスト情報を形態素解析した結果の情報である．名詞比率や MVR，品詞構成率がある [94]．固有表現は，テキスト情報に含まれる固有名詞である．品詞体系に基づく手法と固有表現分類に基づく手法がある．品詞体系に基づく手法には，日本語形態素解析器 MeCab[95][6] で用いられている IPADIC 辞書がある．固有表現分類に基づく手法には，MUC[7] および IREX[8] の規定に基づき拡張された拡張固有表現階層 [96] がある．固有表現クラス分類 [97] で用いられている．評価表現は，テキスト情報の評価を表す表現である [13]．既存研究には，日本語評価極性辞書 [98][99] や単語感情極性対応表 [100]，評価値表現辞書 [101] がある．日本語評価極性辞書は，リスクの見積もり [102] や偏向性を可視化 [103] で用いられている．単語感情極性対応表は，感情推定 [104] の研究で用いられている．評価値表現辞書は，トラブルを表す文の抽出 [105] で用いられている．

[6] MeCab : https://taku910.github.io/mecab/
[7] MUC-6 : https://cs.nyu.edu/cs/faculty/grishman/muc6.html
[8] IREX : https://nlp.cs.nyu.edu/irex/

2.4.2 特徴量選択

CGM やソーシャルメディアを始めとした伝播メディア分析において優れた特徴量のみを選択的に用いる研究に特徴量選択がある．特徴量選択は単変量特徴量選択や，モデルベース特徴量選択，反復特徴量選択などに大別される．単変量特徴量選択には，級間分散を用いる方法などがある [106]．級間分散を用いる方法では，異なる画像が表出されている事例間における分散の大きさで特徴量選択を行う．モデルベース特徴量選択は，アルゴリズムを用いて特徴量選択を行う方法である．反復特徴量選択は，モデルベース特徴量選択のアルゴリズムを用いて，特徴量が特定の特徴量数になるまで，モデルベース特徴量選択を繰り返す方法である．モデルベース特徴量選択においては，多数のアルゴリズムを用いた方法が提案されている．代表的なアルゴリズムに，決定木を用いた特徴量選択 [107] や変数の重要度を用いたランダムフォレストの特徴量選択 [108]，サポートベクターマシンを用いた特徴量選択 [109] などがある．多くの優れたアルゴリズムを持つスパースモデリングを用いる方法などもある [110]．

2.5 既存研究に対する本研究の貢献

前節で述べたように，既存研究における特徴量や特徴量選択を用いて，コミュニケーションから課題解決に有効な少数の特徴量を選択することは，課題解決において非常に有益である．特徴量抽出や特徴量選択には，多様な手法があるが，本研究では，コミュニケーションにおけるメッセージが，複数の高次な特徴量から構成されていることに着目した．

具体例としては，テキストコミュニケーションにおいては，既存研究において，極性反転子[9] 逆接関係を導く表現[10] が存在した場合，以降の関係において，反転する場合があるという指摘がある．そこで，本研究における問題解決では，メッセージを加法的な構成成分に分解する多変量解析に帰着した．既存研究に対する本研究の貢献を下記に示す．

1. 伝播メディアにおけるテキストコミュニケーションに対する幅広い適応性．
2. マス・コミュニケーションも想定した受信者に対する課題解決．
3. 極性反転子や逆接関係を導く表現も包含して評価．

テキストコミュニケーションに対する幅広い適応性においては，第 4 章にて詳述するが，本研究では，既存研究における 31 個，2,071 次元のテキスト情報の特徴量を網羅的に使用した．

マス・コミュニケーションも想定した受信者に対する課題解決においては，1.4 節にて示したが，マス・コミュニケーションにおける 3 要素にプライバシーの要素を含めた 4 要素に基づいた性能評価実験を行う．性能評価実験は，第 4 章および第 5 章にて示す．

テキストコミュニケーションにおける極性反転子や逆接関係については，複数の高次な特徴量であるメッセージ特徴を用いて，課題解決を行う．第 3 章にて示す．

[9] 文末表現における「〜ません」や，英語における「not good」など [13]．
[10] 「しかし」や,「〜だが」など [13]．

第3章　メッセージ特徴と有効性の判別方法

3.1　はじめに

本研究ではメッセージ特徴の抽出に特徴量変換を用いる．特徴量変換には，主成分分析や独立成分分析など，多くの多変量解析手法が提案されている [111]．本研究では，数多い特徴量変換手法の中でも，非負値行列因子分解 (NMF:Non-negative Matrix Factorization) を用いた特徴量変換を用いる [112]．NMF は下記の利点を持ち，他の特徴量変換を用いた次元削減手法と比較し，優れているとされている．

既存研究における NMF の利点 [113]

1. 主成分分析と比較して，非負値制約があるため，結果の解釈が容易．
2. 柔軟なモデルであるため，適用分野に応じて損失関数やアルゴリズムの選択肢が多い．
3. K-means 法や pLSA と関連があり，教師なし学習のモデルとして，理論的な裏付けを持つ．
4. 極めて多様な分野で応用が試みられている．

また，NMF は観測ベクトルを並べた行列が，相関行列や分散共分散行列であることを前提としていない．したがって，行列の各次元の特性に依存せず適用できる．また，NMF はアルゴリズムにおいても，改善が進められている [112]．本章では，まずアルゴリズムである非負値行列因子分解のアルゴリズム概要を示し，NMF における性質と本研究におけるメッセージ特徴の説明を行う．そして，メッセージ特徴が課題に有効な特徴であるかの判別方法の説明を行う．但し，以後の議論におけるメッセージ特徴を評価する伝播メディアのメッセージは，SMCR モデルに基づくコミュニケーションであるものとする．

メッセージ特徴の有効性の判別は分類器，非線形回帰で行う．加えて，本研究では，グラフィカルモデルを用いて，分析課題に応じて選択したメッセージ特徴における特徴量の因果関係推定を行う．

3.2 非負値行列因子分解アルゴリズムと本研究におけるメッセージ特徴

3.2.1 非負値行列因子分解

本章では，メッセージ特徴の分析において，最も重要な役割を果たす非負値行列因子分解である NMF について説明する．NMF は，特徴量変換手法の一つである．本研究においては，メッセージ特徴は評価対象より特徴量を抽出し，特徴量を変換することで生成する．本研究の特徴量変換では，抽出した特徴量を観測ベクトルとし，観測ベクトルを水平方向に並べた行列を観測データ行列と見なす．特徴量は 4.2 節で後述するが，4.2.1 節，4.2.2 節，4.2.3 節の各特徴量を評価対象より抽出する．本研究の特徴量の変換には，NMF を用いる [112]．NMF は観測行列 Y を基底行列 H と係数行列 U の積に分解するアルゴリズムである．観測行列 Y は観測データ行列である．詳細は，文献 [112] にゆずり，ここでは，概略を述べるに留めることにする．NMF の簡略化した分解表現を式 (3.1) に示す．

$$Y \simeq HU \tag{3.1}$$

観測行列 Y は，N 行 K 列の長方行列である．評価対象となる N 個の観測ベクトルを式 (3.2) に示す．

$$y_1, ..., y_N \subset R^{\geq 0,K} \tag{3.2}$$

ここで，$R^{\geq 0,K}$ は，K 次元の非負値ベクトル全体の集合である．評価対象となる観測行列 Y は，観測ベクトルを並べた行列を表し，Y=$[y_1, ..., y_N]$ と表す．NMF では，観測行列 Y の次元数 K よりも，基底数 M を小さく設定することで，特徴量変換が行える．NMF では，$M < min(K, N)$ のとき，観測行列 Y を低ランク行列の積で近似することに相当する．基底行列 H の基底数を $(m = 1, \ldots, M)$ とした際の NMF による観測行列の分解を式 (3.3) に示す．

$$y_n \simeq \Sigma_{m=1}^{M} h_m u_{m,n} (n = 1, ..., N) \tag{3.3}$$

式 (3.3) の y_n は観測行列 Y を表す．また，h_m は基底行列 H の成分である．$u_{m,n}$ は係数行列 U の成分を表す．NMF では，観測ベクトルを最も良く説明する M 個の基底ベクトルおよび重み係数を推定する．したがって，観測ベクトルを並べた行列を 2 つの非負値行列の積に分解する問題であると言える．しかし，行列分解では一般に誤差が発生する．また，行列 H，U は，一意に決まらない．このため，NMF の行列分解は，観測ベクトルを並べた行列である観測行列 Y を H，U の誤差を最小化する行列 H，U を求める最適化問題である．NMF では，乖離度規準を定義し，規準に基づく更新式の反復で最適解を求める．本研究の NMF では，ランダムな非負値で初期化した行列 H，U に式 (3.4) の更新式を収束するまで繰り返し適用する [112]．更新の収束で，行列分解結果の基底行列 H，係数行列 U が得られる．ここで，NMF の乖離度規準は，観測行列 Y の生成プロセスによって異なる．

本研究では，観測行列 Y の生成プロセスに，非負の整数の確率分布である Poisson 分布を仮定した．したがって，本研究における乖離度規準には，一般化 Kullback-Leibler ダイバージェンス [112] を採用した．本研究で用いる一般化 Kullback-Leibler ダイバージェンスに基づいた更新式を式 (3.4) に示す．

$$
\begin{aligned}
h_{k,m} &\leftarrow h_{k,m} \frac{\Sigma_n y_{k,n} u_{m,n} / x_{k,n}}{\Sigma_n u_{m,n}} \\
u_{m,n} &\leftarrow u_{m,n} \frac{\Sigma_k y_{k,n} h_{k,m} / x_{k,n}}{\Sigma_k h_{k,m}}
\end{aligned}
\tag{3.4}
$$

3.2.2 メッセージ特徴

本研究における，メッセージ特徴を説明する．NMF では，$M < min(K, N)$ のとき，観測行列 Y が，低ランク行列の積で近似することに相当する．したがって，観測行列 Y は，基底行列 H と係数行列 U の線形結合で表現できる．線形結合の例を式 (3.5) に示す．

$$
h_1 \begin{pmatrix} u_{1,n} \\ u_{2,n} \\ \vdots \\ u_{M,n} \end{pmatrix} + \cdots h_M \begin{pmatrix} u_{1,n} \\ u_{2,n} \\ \vdots \\ u_{M,n} \end{pmatrix}
\tag{3.5}
$$

式 (3.5) の係数行列 U は，の成分 $u_{M,n}$ は文書 n の基底 M への重みである．また，基底行列 H の成分 h_M は，基底 M への各特徴量の寄与率の集合である．ここで，基底 M は，観測行列 Y の特徴量の共起成分がグルーピングされた結果である．本研究では，グルーピングされた基底を，伝播メディアの課題に応じて評価し，基底を選択する．本研究における基底はメッセージ特徴に相当する．本研究では，課題解決に有効な基底であるメッセージ特徴を評価した．本研究のメッセージ特徴の選択では，2 種類のメッセージ特徴の選択手法を用いる．2 種類のメッセージ特徴の選択手法は，メッセージ特徴の評価方法が異なる．

本研究では，第 4 章で，差異に基づくメッセージ特徴の特徴選択手法とその性能評価実験を論じる．また，第 5 章で，回帰に基づくメッセージ特徴の特徴選択手法とその性能評価実験を論じる．

3.3 分類アルゴリズムを用いたメッセージ特徴の有効性判別

3.3.1 Ada Boost

AdaBoost は，ブースティング方式の分類手法である．ブースティングは，精度が低い分類器である弱分類器 (Weak classifier) を複数組み合わせることで，高精度の分類器 (Strong classifier) を構築するアルゴリズムである．AdaBoost は，最初はすべての訓練データに等しい重みを与える．その後，各ラウンドで，分類を誤った事例の重みを指数的に増やし，より難しい事例を集中して学習する．そして，分類誤り率から，適応的 (adaptive) に仮説に対する重みと次のラウンドの訓練データに対する重みを決定することから AdaBoost(adaptive boosting) と呼ばれる [114]．AdaBoost では分類規則の誤り確率が小さいほど，重みが大きくなるように設定されている．また，多くの理論的検証と実験的実証から有効性が示されているアルゴリズムである [115]．詳細は，文献 [114][115] にゆずり，ここでは，概略を述べるに留めることにする．

まず，訓練データである $(x_1, y_1)...(x_m, y_m)$ を獲得する．ここで，x_i は入力ベクトル，y_i は分類ラベルであり，$y_i \in \{0, 1\}$ である．訓練データの確率分布を式 (3.6) で初期化する．

$$D_1(i) = \frac{1}{m} \tag{3.6}$$

その後，式 (3.7) から式 (3.9) の処理を繰り返す．繰り返し処理では，第一に式 (3.7) で誤り率 ε_t を得る．第二に，重み更新係数 α_t を得る．第三に，重みの分布 D_t を更新する．更新で用いる Z_t は正規化定数である．

$$\varepsilon_t = \sum_{i:h_t(x_i) \neq y_i)} D_t(i) \tag{3.7}$$

$$\alpha_t = \frac{1}{2} log(\frac{1 - \varepsilon_t}{\varepsilon_t}) \tag{3.8}$$

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-a_t} & if \quad y_i = h_t(x_i) \\ e^{a_t} & if \quad y_i \neq h_t(x_i) \end{cases} \tag{3.9}$$

繰り返し処理後，獲得した誤り率 ε_t と重み更新係数 α_t を用いて，式 (3.10) で，強分類器である H を獲得する．

$$H(x) = sign\{\Sigma_t \alpha_t h_t(x)\} \tag{3.10}$$

3.3.2 Random Forests

Randam Forests は，複数の決定木 (decision tree) を用いた分類手法である [116]．決定木は，あるデータ集合とその属性で木構造を構成し，識別ルールを構築する手法である．詳細は，文献 [116] にゆずり，ここでは，概略を述べるに留めることにする．Randam Forests と決定木では，ジニ係数を用いて不純度を算出し，分割の評価基準としている．ジニ係数の算出式を式 (3.11) で示す．

$$Gini = \sum_{i=1}^{K} P(C_i)(1 - P(C_i)) = 1 - \sum_{i=1}^{K} P^2(C_i) \tag{3.11}$$

K はクラス数，$P(C_i)$ はノード t に分岐するサンプルがクラス i に属する確率である．したがって，$\sum_{i=1}^{K} P(C_i)(1 - P(C_i))$ はノード t における誤り率である．Randam Forests では，データ集合 I からランダムサンプリングで S 個のサンプルを抽出し，決定木を構築する．Randam Forests の決定木は二分木で構築される．Randam Forests アルゴリズムの学習フローを下記に示す．

Randam Forests アルゴリズムの学習フロー [116][117]

(STEP1)
データ集合 I から，ランダムサンプリングで S 個のサブセットを生成．

(STEP2)
S 本の決定木が学習完了するまで (STEP3) を繰り返す．

(STEP3)

1. 属性 (特徴量) をランダムに選択．
2. ジニ係数に基づき，選択した変数の中から，最適に分割する属性 (特徴量) と分割点を決定．
3. 末端ノードに達する． OR 指定した深さの階層に達する．
4. 決定木の学習完了を確認．

(STEP4)
S 本の決定木の学習完了を確認し，終了．

S 本の決定木の学習完了で Randam Forests モデルが構築される．Randam Forests を分類器として，分類を行う場合，構築された S 本の決定木で，各決定木の分類結果を集計し，結果として出力する．

3.3.3 Multi-Layer Perceptron

MLP(Multi-Layer Perceptron) は，入力信号を，出力信号に変換する (無記憶) 非線形の神経回路網モデル (ニューラルネットワークモデル) である [118][119]．また，MLP は多層パーセプトロンとも呼ばれ，入力層，出力層およびいくつかの中間層 (hidden 層) からなり，入力から出力の方向にいくつかの層間の結合がある．MLP に関して簡単な説明を行う．詳細は，文献 [118][119] にゆずり，ここでは，概略を述べるに留めることにする．入力信号 x に対する出力信号 y，ガウス雑音 n とした際の MLP を式 (3.12) に示す [119]．

$$y = f(x, \theta) + n \tag{3.12}$$

式 (3.12) の θ は MLP のパラメータをすべてまとめたベクトルである．MLP は多層から構成されている．入力ベクトルを w，重みベクトルを w_i とした際の，MLP を式 (3.13) で示す．ここで，隠れ層の変数である i は，$i = (1...h)$ である．

$$f(x, \theta) = \sum_{t=1}^{b} v_i \varphi(w_i \bullet x) \tag{3.13}$$

式 (3.13) では，第 i 番目の隠れ層素子は，非線形関数である $\varphi(w_i \bullet x)$ を出力する．$w_i \bullet x$ は入力ベクトル x と重みベクトル w_i を用いた内積である．また，$\varphi(u)$ は飽和型関数 (シグモイド関数) である．最後の出力素子では，飽和型関数 $\varphi(w_i \bullet x)$ の結果を重み v_i で総和する．θ は $(w_i, \cdots, w_h; v_i, \cdots, v_h)$ であり，ニューラルネットワークのパラメータをすべてまとめたベクトルである．ここで，最終出力結果がガウス雑音 n で乱された場合の結果は，入出力特性は条件付き確率分布である式 (3.14) となる．

$$p(y|x; \theta) = cexp\{ -\frac{1}{2}(y - f(x, \theta))^2 \} \tag{3.14}$$

MLP などニューラルネットワークを分類器として，分類を行う場合，出力信号 y に相当する出力層の出力を分類クラスに対応させ，分類を行う [120]．

3.3.4 K-Nearest Neighbours(K-NN)

K-Nearest Neighbours 判別 (K-NN 判別) は正しく分類が行われている既存のデータから，判別を行いたい個体に最も「近い」個体 (最近傍点) を k 個選び出し，それらの個体が最も多く属しているクラスに当該固体を分類する分類器である [121]．K-NN 判別では，最近傍点の個数と，個体までの距離に用いる距離測度がパラメータとなる．K-NN 判別では，最近傍点の数である K でモデル分類器の複雑性が決定される．複雑度が少ない場合は適合不足，複雑度が高い場合は過剰適合となる．多くの最近傍点を用いた場合は，なめらかな決定境界となる．したがって，複雑度の低い単純なモデルとなる．一方で，最近傍点が少ない場合は複雑度の高いモデルとなる．本研究では，最近傍点の数は $K = 5$ である．距離測度は，判別を行いたい個体から最近傍点までの距離規準である．本研究の K-NN 判別では，ユークリッド距離を用いた．

3.3.5 適合率・再現率・F-measure

適合率は不正解データを正解データと判定しないようにする評価指標である．再現率は，分類器がすべての正解データを判定する評価指標である．また F-measure は適合率と再現率の調和平均値である．評価指標を図 3.1 に示す．適合率，再現率，F-measure を式 (3.15)，式 (3.16)，式 (3.17) に示す．また，偽陽性率を式 (3.18) に示す．偽陽性率は，陰性サンプル個数に対する偽陽性数の割合である．

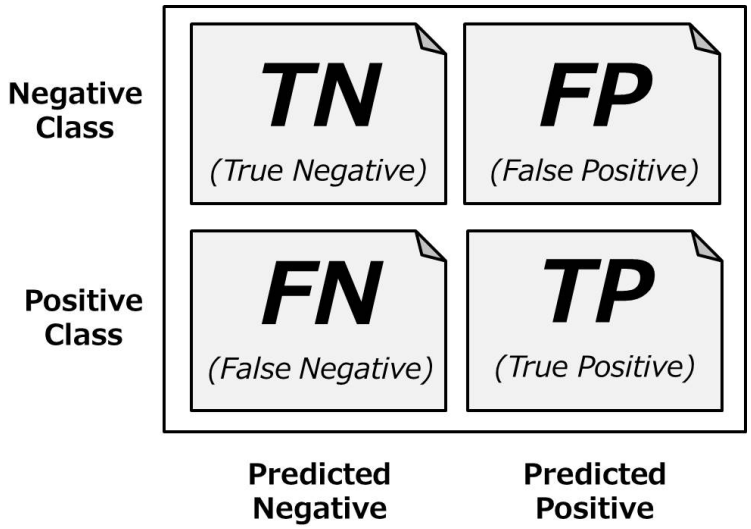

図 3.1: Contingency Table

$$Precision = \frac{TP}{TP + FP} \tag{3.15}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.16}$$

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.17}$$

$$FPR = \frac{FP}{FP + TN} \tag{3.18}$$

3.4 非線形回帰を用いたメッセージ特徴の有効性判別

3.4.1 サポートベクタ回帰モデル

本研究では，提案手法の評価に非線形回帰手法の一つであるサポートベクター回帰モデル (SVR : Support Vector Regression)[122] 用いる．SVR は，入力 $x_i \in R^p (i = 1, 2..., l)$ を特徴空間へ非線形写像し，特徴空間で線形回帰を行うモデルである．また，SVR は汎化能力が高い回帰モデルであることが知られている [123]．詳細は，文献 [122][123][124] にゆずり，ここでは，概略を述べるに留めることにする．

SVR の回帰関数を式 (5.1) に，回帰関数で用いるカーネル関数を式 (5.2) に示す．本研究では，カーネル関数に RBF カーネル [124] を用いる．

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x) + bias \tag{3.19}$$

$$k(x_i, x) = \exp(\parallel x_i - x \parallel^2 / 2\sigma^2) \tag{3.20}$$

$K(x_i, x)$ は入力 x_i を特徴空間へ写像するカーネル関数である．α_i,α_i^*,$bias$ などの詳細は文献 [124] などを参照していただきたい．

3.4.2 予測誤差

予測誤差には，MAE(Mean Absolute Error) および RMSE(Root Mean Squared Error) を用いた．MAE を式 (3.21) に RMSE を式 (3.22) に示す．n は予測対象のデータ数，y_i は実績値，$\widehat{y_i}$ は予測値である．

$$MAE = \frac{1}{n} \Sigma_{i=1}^{n} |y_i - \widehat{y_i}| \tag{3.21}$$

$$RMSE = \sqrt{\frac{1}{n} \Sigma_{i=1}^{n} (y_i - \widehat{y_i})^2} \tag{3.22}$$

3.5 グラフィカルモデルを用いたメッセージ特徴における因果関係推定

3.5.1 ベイジアンネットワーク

本研究では，メッセージ特徴から得られる特徴量の “因果関係” の評価でベイジアンネットワーク (Bayesian Network) を用いる [1][125]．ベイジアンネットワークは，ベイズの定理 (Bayes rule) に基づいたモデルである．ベイジアンネットワークの例を図 3.2 に示す．

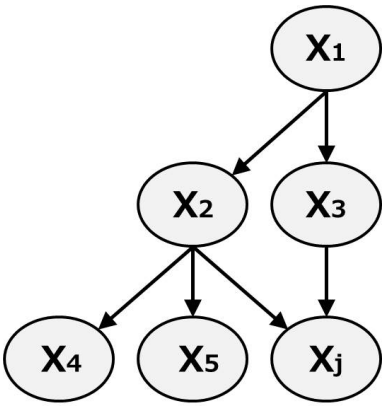

図 3.2: Bayesian Network

図 3.2 の例では $X1$ が親ノードであり原因，$X2$ や $X3$ が子ノードであり結果である．$X2$ や $X3$ のように，子ノードは，他の子ノードである Xj の親ノードにもなる．

ここで，事象 $P(X_1)$ を原因，事象 $P(X_j)$ が結果である場合，原因によって結果が起きる確率は $P(X_j|X_1)$ で表される．ベイズの定理では，この $P(X_j|X_1)$ に加えて，$P(X_1)$，$P(X_j)$ が明らかな場合，結果に対する原因の確率である $P(X_1|X_j)$ を推定することができる．原因の事象間の関係性を多数のノードに拡大したモデルがベイジアンネットワークである．ベイジアンネットワークは，各変数は確率変数，ノード間の矢印は因果関係，条件付き確率で定量化されている有向非循環グラフで表される．ベイジアンネットワークでは，結果である子ノードを X_j とした際，原因 $P_\alpha(X_j)$ は，親ノードの集合 $(x_1^j, \cdots, x_i^j)$ である．したがって，依存関係は $P(X_j|P_\alpha(X_j))$ と表すことができる．このため，すべての確率変数の同時確率分布は式 (3.23) と表すことができる．

$$P(X_1, \cdots, X_n) = \prod_{j=1}^{n} P(X_j|P_\alpha(X_j)) \tag{3.23}$$

すべての確率変数の同時確率分布が明らかな場合，確率変数の依存関係は各子ノードとその親ノードの間にリンクを張って，ベイジアンネットワークのグラフ構造で表される．したがって，構築したグラフ構造から変数間の依存関係が明らかになり，特性のノードから，原因ノードを推定できる．

第4章　差異に基づくメッセージ特徴の選択手法とその性能評価

4.1　はじめに

近年では，科学技術に基づいた場所に依存しないコミュニケーションの増加で，情報受信チャネルと情報取捨判断の機会が増加した．発信された情報は，受信者に伝達されるかは紙一重の状況が頻繁化していると言える．メッセージの受信時に不要と判断された場合，メッセージの内容が閲覧されることはない．受信者が有益でないと判断した伝播メディアの情報発信も同様である．ゆえに，情報の発信者は，受信者が有益と判断するメッセージの内容や，伝播メディアなど，メッセージ特徴の選択において，重要視するコンテンツの要素を明らかにしたいという要望が高まっている．差異に基づくメッセージ特徴では，利用者が有益と判断するコンテンツのメッセージ特徴を評価対象とするメッセージ集合の特性に基づいて評価する．

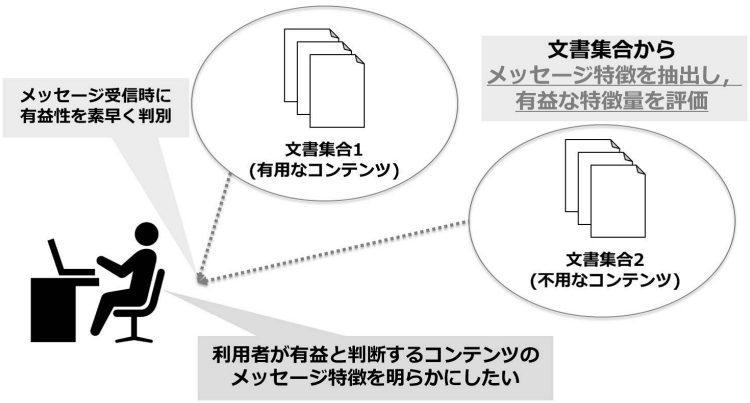

図 4.1: 評価集合の差異における問題

差異に基づくメッセージ特徴は，特徴量の抽出，メッセージ特徴の抽出，メッセージ特徴の選択で構成されている．差異に基づくメッセージ特徴の評価実験では，既存研究の特徴量抽出手法を用いて，テキスト情報の特徴量を抽出する．特徴量の変換では，非負値行列因子分解 (NMF) を用いて，メッセージ特徴を抽出する．そして，提案手法を用いて，差異に基づくメッセージ特徴を抽出する．メッセージ特徴は複数の特徴量からなる特徴量の集合である．

以下，4.2 節にて，評価実験で用いるテキスト情報とその抽出方法を示す．メッセージ特徴は，観測行列に対し，4.3 節で示す NMF を適用した結果の基底である．メッセージ特徴を評価する方法を示し，4.2.4 節にて実装方法を示す．そして，4.4 節および 4.5 節にて，提案手法の評価実験を示す．

4.2 テキストコミュニケーションメッセージ

本研究では，メッセージから，特徴量を抽出し，非負値行列因子分解 (NMF) を適用することで高次なメッセージ特徴を抽出する．本研究では，特徴量抽出において，複雑性を内包するため，既存研究から，多くの特徴量を抽出する．本研究の評価実験では，テキスト情報の特徴量を用いた．テキスト情報の特徴量は，表層情報，トピックである話題，語種，基本語，意味属性，言語表現，文末表現，品詞，固有表現，評価表現の 10 種類，31 個，2,071 次元の特徴量である．ここでは，既存研究の特徴量抽出方法を，抽出方法ごとに表層の特徴，トピックの特徴，意味情報の特徴の 3 グループとして，詳述する．

4.2.1 表層の特徴抽出

テキスト情報の表層の特徴を表 4.1 に示す．表 4.1 の 8 種類の特徴量でコンテンツである文面の表層の特徴を捉えることとした．本研究の文字種には，ひらがな[1]，カタカナ[2]，漢字[3]，アルファベット[4]，数字[5]，半角記号[6]，空白記号[7]，全角記号[8]を採用した．

[1]仮名の一つ．平安初期に成立した音節文字．漢字の草書体から作られた草仮名をさらに簡略したもの．

[2]「ア (阿)」「イ (伊)」「ウ (宇)」のように，多く漢字の一部をとって作り出された表音文字．平安初期作られた.「かた」は完全ではないの意．ひらがなに対応する体系を持ち，外来語，動植物名などに用いられる．

[3]中国で作られ，日本などでも使われている表語文字．最古の漢字は殷の甲骨文字．一字一音節で一語を表す．日本で作られた国字を含めて言う．

[4]セム語系の文字に起源し，英語・ギリシャ語など現代世界の諸言語で用いられている文字．ローマ字 26 文字を言う．

[5]数を表す文字．数を記録することは文字成立以前から，樹幹や岩にきざみつけることなどによって行われており，古代エジプトの象形文字やバビロニアの楔形数字が作られた．

[6]記号は一定の事象や内容を代理・代行して指し示すはたらきをもつ知覚可能な対象．半角文字は正方形の和文活字一字を半分にした大きさの文字．

[7]文章入力字にスペースキーを押下することで，入力される文字コード．文字と文字の間隔を空けるなどの目的で入力する．

[8]正方形の和文活字一字分の大きさ．2 バイト文字．

*Unicode 10.0 Character Code Charts : https://www.unicode.org/charts/

表 4.1: 表層情報

特徴量名	次元数	値の定義/例
文の数	1	1 文書中の [。][？][！] の合計頻度
読点	1	1 文書中の [、] の頻度
句点	1	1 文書中の [。]
文の長さ	1	1 文書中の総バイト数
文字種 (頻度)	8	各文字種の出現頻度*
文字種 (比率)	8	各文字種の比率*
読点間距離	1	文献 [73]
漢字含有率	1	文献 [73]

4.2.2 トピックの特徴抽出

ここでのトピックはテキスト情報が言及する話題である．話題を抽出する手法では，テキストは複数の単語 (語彙) が集まったもの (Bag-of-words) として捉える．そして，類似した語彙の集合で話題 (トピック) が形成されるとしている．本研究では話題抽出で用いられる基本的なアルゴリズムである LSI(Latent Semantic Indexing) および LDA(Latent Dirichlet Allocation)[22] を用いる．各アルゴリズムに関して，簡単な説明を行う．

LSI は，文書集合を行列 N として，低ランク行列の積 $U^{\mathrm{T}}H$ にそれぞれの行列の要素を二乗し，総和をとったものが最小になるように行列分解する手法である．行列 N を文書番号 D と語彙 V とした際の，LSI を式 (4.1) で示す．また，文書集合の行列 N に LSI を適用した図を図 4.2 に示す．

$$\|N - U^{\mathrm{T}}H\|^2 FRO = \sum_{d=1}^{D}\sum_{v=1}^{V}(N_{dv} - u_d^{\mathrm{T}}h_v)^2 \tag{4.1}$$

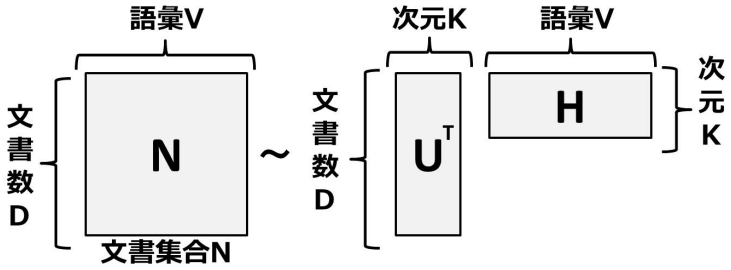

図 4.2: LSI(Latent Semantic Indexing)

$U = (u_1, \cdots, u_D)$ は重み付き係数行列であり，K 行 D 列の実数行列である．$H = (h_1, \cdots, h_V)$ は語彙行列であり，K 行 V 列の実数行列である．ここで，K は低ランク行列の次元数である．LSI の行列分解には，特異値分解が用いられている．4.2.4 節で後述するが本研究の LSI の実装は，分かち書き[9]されたテキスト情報に，Gensim[126][10]を適用する．したがって，特異値分解は，乱択特異値分解 [127][11]を用いる．

LDA は，文書集合の行列 N の文書が低ランク行列 θ と ϕ をパラメータとして持つカテゴリ分布 Λ から生成されると仮定する手法である．全体集合を W，文書数を $d = (1, \cdots, D)$，トピックを $k = (1, \cdots, K)$ とした際の，文書 w_d の生成確率を式 (4.2) で示す．また，文書集合の行列 N に LDA を適用した図を図 4.3 に示す．

$$p(w_d|\theta_d, \Phi) = \prod_{n=1}^{N_d} \Sigma_{k=1}^{K} p(k|\theta_d) p(w_{dn}|\phi_k) \tag{4.2}$$

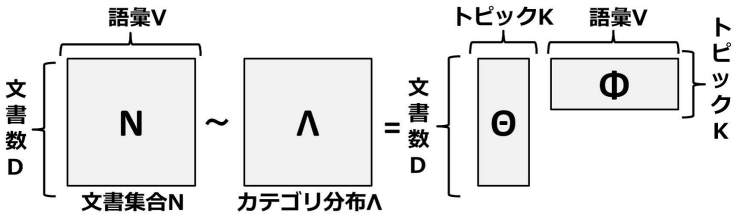

図 4.3: LDA(Latent Dirichlet Allocation)

パラメータ θ_d はトピック分布，パラメータ Φ は単語分布集合である．また，N_d は文書 d に含まれる単語数，w_{dn} は文書 d の n 番目の単語，ϕ_k はトピック k の単語分布を示す．本研究における LDA の実装に関しては，LSI と同様に Gensim[126] を用いた．よって，各文書のトピック分布 θ_d の推定は，変分ベイズ法 [128] を用いる．

ここで，文書集合である行列 N は 1 行が 1 文書，各列は語彙に相当し，要素は語彙 V の出現頻度である．LSI の場合，次元 K が話題であり，文書に対する次元 K の重みである U の要素が特徴量に相当する．LDA の場合，トピック K が話題であり，文書に対するトピック K の割り当て確率であるトピック分布 θ の要素が話題に相当する．アルゴリズムでは，LSI の次元数や LDA のトピック数は任意の数の設定が行える．LSI の次元数や LDA のトピック数である K は，任意で値の設定を行う．LSI の K の最適なパラメータは 300 から 500 の範囲が提案されている [129]．本研究では K=300 を設定した．LDA に関しても同様に K=300 を設定した．よって，特徴量の次元数は 600 である．話題の特徴量を表 4.2 に示す．

[9] 一定の規則にしたがって，区切られた文章を分かち書きと呼ぶ．
[10] gensim : https://radimrehurek.com/gensim/
[11] Blei Lab : https://github.com/blei-lab

表 4.2: 話題

特徴量名	次元数	値の定義
潜在意味解析 (LSI)[77]	300	文書の重み付き係数 u_D
トピックモデル (LDA)[78]	300	文書のトピック分布 θ_d

LSI では，低ランク行列分解された重み付き係数行列 U の値によって，文書 D における話題 K の話題の重み付きが明らかになる．よって，LSI の特徴量には，各文書の K の重み付き係数を用いている．LDA では，各文書における θ_d を推定することで，話題の割合が明らかになる．よって，LDA の特徴量には，各文書の θ_d を用いている．表 4.2 の特徴量の値は実数である．

4.2.3 意味情報の特徴抽出

意味情報は，個々の研究で言語や形態論に基づき，異なる特徴が用いられている [81]．本研究では，学術目的の言語資源として，従来研究で作成された日本語の辞書のうち，言語資源の利用事例や応用研究がある既存辞書 21 個を選定した．本研究で用いる 21 個の辞書を表 4.3 に示す．

表 4.3 のうち，項番 15 の拡張固有表現の辞書は，関根の拡張固有表現階層の定義に基づき，Wikipedia の見出し語に対し，固有表現クラスを付与した辞書である．また，項番 18-20 で用いている単語感情極性対応表では，日本語の辞書の見出し語と読みを用いた．加えて，表 4.3 の項番 1-7 の辞書は調査研究目的に作成された辞書に基づいている．項番 1，3，6 の辞書は意味分類体語彙表，項番 2，5 の辞書は日本語教育基本語彙，項番 4，7 の辞書は分類項目一覧表である．本研究では，下記に記載の基準で，辞書の加工および見出し語を選定した．

1. **見出し語の加工**
   空白記号,「-」「0」「など」を除去

2. **対象外の見出し語**
   「」「-」「〔」「→」「/」「(」「」「その他」
   を含む語

本研究における意味情報は，表 4.3 で示した項番 1 の語種，項番 2-4 の基本語，項番 5-7 意味属性，項番 8 の言語表現，項番 9 の文末表現，項番 11-14 の品詞，項番 15 の固有表現，項番 16-21 評価表現の 8 種類を評価に用いる．

表 4.3: 意味情報

項番	特徴量名	次元数	語数	値の定義/付与タグの例	文献/脚注
1	語種	7	6519	和語，漢語，外来語	†,[83]
2	基本語 (1)	2	697	基本語二千に選定	†,[84]
3	基本語 (2)	6	6519	外国語学習基本語彙	†,[84]
4	基本語 (3)	2	424	基本語二千・六千に選定	†,[84]
5	意味分類コード (1)	233	697	体の類抽象的関係	†,[86]
6	意味分類コード (2)	487	6519	(同上)	†,[86]
7	意味分類コード (3)	307	424	(同上)	†,[86]
8	機能表現	122	29262	O，B-判断	‡,[88]
9	文末モダリティ	32	32	可能性，表出	[90]
10	質問文末表現	38	38	質問，回答	[91]
11	IPA 品詞	14	-	連体詞，接頭詞，名詞	§,[95]
12	名詞比率	1	-	実数 (算出 文献 [94])	[94]
13	MVR	1	-	実数 (算出 文献 [94])	[94]
14	固有名詞	4	-	実数 (算出 文献 [95])	§,[95]
15	拡張固有表現	132	18075	数値表現	¶,[96]
	評価極性情報				
16	(用言編)	4	5280	ネガ (経験)	‡,[98]
17	(名詞編)	51	13314	～になる (状態) 客観	‡,[99]
	感情極性		110250		‖,[100]
18	(頻度)	2	-	ポジティブ・ネガティブ	[13]
19	(比率)	2	-	実数 (算出 [13])	
20	(平均値)	1	-	実数 (算出 [13])	
21	評価値表現	1	5234	-	**,[101]

†『日本語教育のための基本語彙調査』データ https://mmsrv.ninjal.ac.jp/bvjsl84/
‡機能表現タグ付与コーパス，日本語評価極性辞書 https://www.cl.ecei.tohoku.ac.jp/index.php
§ipadic version 2.7.0 ユーザーズマニュアル https://chasen.naist.jp/snapshot/ipadic/ipadic/doc/ipadic-ja.pdf
¶NAIST Japanese ENE Dictionary on Wikipedia https://github.com/masayu-a/NAIST-JENE
‖単語感情極性対応表 https://www.lr.pi.titech.ac.jp/takamura/pndic_ja.html
**評価値表現辞書 https://www.syncha.org/evaluative_expressions.html

4.2.4 実装方法

本研究の実装は，プログラム言語 Python [12]を用いた．単語分割，品詞判定は，MeCab[95] を用いた．アルゴリズムの実装は，Gensim[126]，scikit-learn[13] [14]を用いた．テキスト情報の前処理は，文字コードを UTF-8 に変換，空白記号・改行コードの除去を実施した．バイト列の欠損行は対象外とした．テキスト情報の等価な文字は Normalization Form KC (NFKC) [15]で正規化を実施した．また，形態素解析では，活用形を標準形に変換し，1 文字単語などを削除する前処理を実施した．観測行列 Y 作成は，Pandas[16] および Numpy[17]を用いた．

本研究で話題の特徴量抽出に用いる LSI の次元数や LDA のトピック数では，任意で値の設定を行う．LSI の K の最適値は 300 から 500 の範囲が提案されている [129]．本研究では K=300 を設定した．LDA に関しても同様に K=300 を設定した．

提案手法で用いる NMF の観測行列 Y は，$(j = 1, \ldots, K)$ から構成される N 行 K 列の長方行列である．ここで，特徴量の次元数 K は，表層情報の次元数 22，アルゴリズムの次元数 600，辞書の次元数 1,451 の合計値 2,071 である．後述するが，5.3 節の実験は辞書にプライバシー侵害に関連のある既存研究の辞書を用いた．したがって，5.3 節の実験の観測行列は合計値 2,073 である．特徴量の値は 0 から 1 の間に正規化し，各変数の計測尺度の違いを考慮するため，L1 正則化による制約補正を実施した

ベイジアンネットワークのモデル構築では，山登り法 (Hill Climbing)，グラフ構造の評価に用いる評価基準には，AIC(Akaike's Information Criterion) を用いて構築を行った．

[12]Welcome to Python.org : https://www.python.org
[13]scikit-learn : https://scikit-learn.org/stable/
[14]scikit-learn : https://github.com/scikit-learn
[15]Unicode Technical Reports : https://unicode.org/reports/
[16]Pandas : https://pandas.pydata.org/
[17]NumPy : https://www.numpy.org

4.3 差異に基づくメッセージ特徴の選択手法

差異に基づくメッセージ特徴の抽出ではグレゴリー・ベイトソン[18] の情報の定義に着目している．ベイトソンは情報を「“違い” を生む “違い”」であると定義している [130]．“違い” を土地と地図の差異に例えた場合，土地には，高低，建造物，人口の分布など，多様な要素がある [131]．また，土地が違えば地図も異なる．地図には，高低図，街の配置図，人口分布図などがある．ここで，地図は土地の特定の要素を選択した結果であり，他の地図との差異に基づく「“違い” を生む “違い”」である．本研究における特徴量変換は，3.2.1 節にて示した非負値行列因子分解である NMF にて行う．NMF の行列分解では，係数行列 U の非負性の制約で，係数行列 U の要素がスパースになる傾向がある [112]．

係数行列 U の要素がスパースであることは，成分である $u_{M,i}$ が文書 i のコンテンツによって，重みがある基底と重みが 0 の基底の差が明瞭であることに相当する．したがって，特性に基づく 2 種類の評価対象の集合がある場合，評価対象の集合の一方でのみ基底の重みが高い基底は，評価対象の集合の特性を表す基底である．

基底の評価が行えた場合，寄与率を用いて目標変数に影響の大きい特徴量の評価が行える．観測行列 Y に NMF を適用した結果，質問記事の各基底の重み付きは係数行列 U で表される．そこで本研究では，目標変数に基づき，係数行列 U の行ベクトルにラベル付けする．そして，ラベル付けに基づき，係数行列 U を 2 種類の評価対象の集合に分割する．本研究では，目標変数は正規分布にしたがうと仮定し，平均値以上を 1，平均値未満を 0 としてラベル付けを行う．そして，対応する係数行列 U の行ベクトルをベクトル集合 L_1，L_0 に分割する．

目標変数の集合 X を式 (4.3) に，ラベル付けの方法を式 (4.4) に，ラベル付けに基づく係数行列 U の分割の方法を式 (4.5) に示す．式 (4.4) は，x_i が平均値以上の場合に 1，平均値未満の場合に 0 に，ラベル付けされることを表している．また，x_i は文書 i の閲覧数，u_i は文書 i の係数行列 U の行ベクトルである．

$$X = \left[\begin{array}{cccc} x_1, x_2 & ... & x_N \end{array} \right]^T \tag{4.3}$$

$$x_i = \begin{cases} \frac{\Sigma_{j=1}^{N} x_j}{N} \leq x_i, x_i \in Set & 1 \\ \frac{\Sigma_{j=1}^{N} x_j}{N} > x_i, x_i \in Set & 0 \end{cases} \tag{4.4}$$

$$u_i = \begin{cases} x_i = 1, \quad u_i \in Set & L_1 \\ x_i = 0, \quad u_i \in Set & L_0 \end{cases} \tag{4.5}$$

[18] アメリカの人類学者 (1904-1980)，民族誌「ナベン」「バリ」や人間関係論におけるダブル・バインド論など，文化とパーソナリティ，コミュニケーションに関する理論で，学際的な功績を残した．

ベクトル集合 L_1 および L_0 を基に，集合における基底 M の係数の平均値を算出し，ベクトル集合における基底の重みを算出する．ベクトル集合はそれぞれ目標変数が多い集合，目標変数が少ない集合である．ここで，それぞれの集合で基底の重みが大きい場合は，重みが大きい重要な基底だが，特性を表す基底ではない．評価対象の集合の一方でのみ基底の重みが高い基底は，集合の特性を表す基底である．そこで，各集合の基底 M の係数の重みの差で，評価対象の集合における基底 M の特性を評価する．評価対象の集合における各基底の重みの算出式を式 (4.6) 示す．ラベル付けされたベクトル集合は L_j であり，j は 1 もしくは 0 である．$L_{j,M}$ は L_j における基底 M の係数の重みである．

$$L_{j,M} = \frac{\Sigma_{i=1}^{N} u_{M,i}}{N} \tag{4.6}$$

式 (4.6) における $u_{M,i}$ はベクトル集合 L_j にラベル付けされた行ベクトル u_i の基底 M への重みである．式 (4.6) の結果を基にした基底の特性を表す集合を S とし，本研究の基底 M の評価方法を式 (4.7) に示す．

$$S_M = L_{1,M} - L_{0,M} \tag{4.7}$$

式 (4.7) の S_M の値が最も大きい基底が，特性に基づく変換特徴量であり，差異に基づくメッセージ特徴である．

4.4 - 実験 1-1 - 伝播メディアにおける話題性の課題への応用

4.4.1 概要

“実験 1-1” では，質問記事の閲覧数と既存研究の特徴量で話題性のメッセージ特徴を明らかにする．CGM やソーシャルメディアなどのオンラインコミュニティは誰もが気軽に閲覧・質問・回答が行える．したがって，情報収集や気楽なコミュニケーション，時間をかけた議論など目的に応じた多くの利点がある．また，商品やサービスの利用者が，内容や使い勝手などの情報交換も盛んに行われている．そこで，共感された価値ある情報は，情報伝播や購買行動に影響を与えることが大きい．共感に基づいた生活者消費行動モデルは，SIPS（Sympathize Identify Participate Share & Spread）[17] [19] [20] と呼ばれる．一方で，クレームや批判的なコメントも急激に広がる．このため，企業は閲覧数の増加傾向がある，話題性やクレームがある質問記事を検知・予測し，早急に対処することが欠かせない．

質問記事における話題性の多寡を推定することを目的に差異に基づくメッセージ特徴を用いて，質問記事の評価手法を提案する．本研究における質問記事の話題性は多くの利用者が，共感・関心を持って質問記事を閲覧する状態とする．このため，質問記事の閲覧数に基づいて評価する．閲覧数に基づいて評価することで，質問記事のコンテンツで，閲覧数の異常や変化の兆し，投稿後に話題になる質問記事を検知することが期待できる．

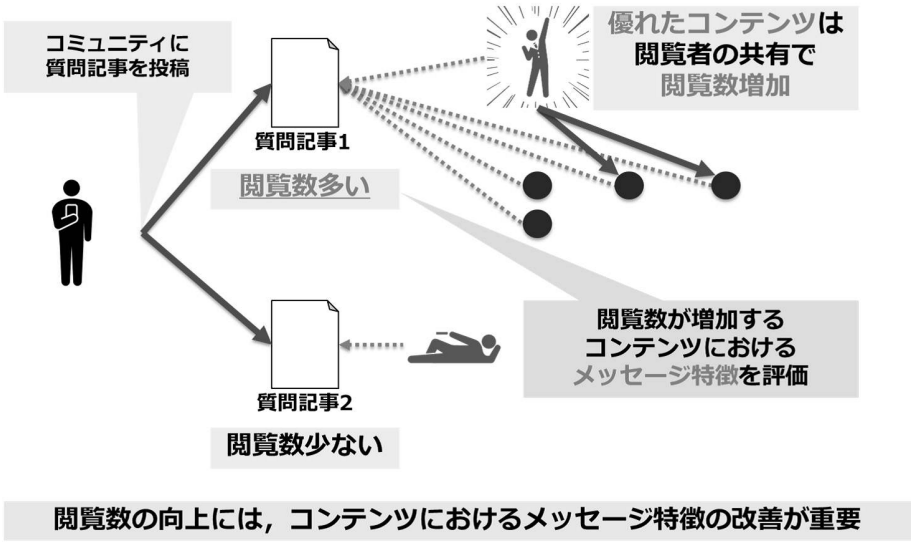

図 4.4: 研究背景 (実験 1-1)

[19] SIPS では，生活者消費行動を S(Sympathize : 共感する)，I(Identify : 確認する)，P(Participate : 参加する)，S(Share & Spread : 共有・拡散する) とモデル化している．SIPS モデルでは，共感に次ぐ重要な要素は，参加してもらうことである．参加はエバンジェリスト (伝道者)，ロイヤルカスタマー (支援者)，ファン (応援者)，パーティシパント (参加者) など，企業やブランドの生涯顧客価値を高めていく過程と重なっている．

[20] SIPS : https://www.dentsu.co.jp/sips/index.html

“実験 1-1” における評価実験手順の概要を図 4.5 に示す．図 4.5 の (1) から (4) が，“実験 1-1” のメッセージ特徴の抽出手順を示すステップである．

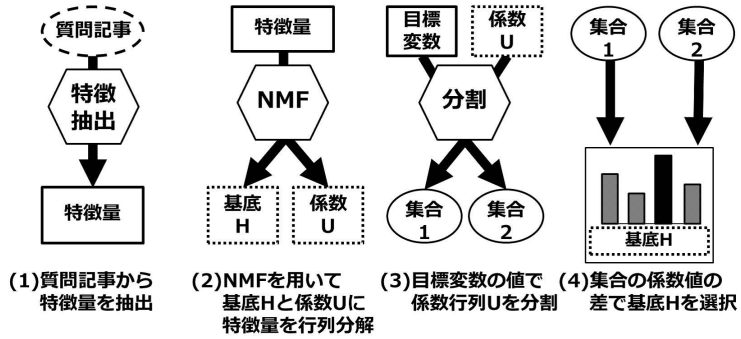

図 4.5: 実験 1-1 メッセージ特徴抽出手法

図 4.5(1) の特徴量の抽出では 4.2.1 節，4.2.2 節，4.2.3 節に示した特徴量を抽出している．抽出する特徴量は既存研究の特徴量を組み合わせた 31 個である．また，各特徴量を水平方向に連結した際の次元数は 2,071 次元である．そして (2) から (4) が，“実験 1-1” のメッセージ特徴を評価する手順であり，差異に基づくメッセージ特徴の抽出に相当する．(2) では NMF を用いて行列分解を行い，特徴量の変換を行う．(2) および (4) のメッセージ特徴における基底選択では，4.3 節の差異による基底選択を用いた．“実験 1-1” の目標変数には，質問記事の閲覧数を用いた．“実験 1-1” は，2.1 節で示した従来研究と比較し，テキスト情報の特徴量を網羅的に抽出している．また，話題性の評価の従来研究で，NMF による行列分解で特徴量変換し，網羅的なテキスト情報から基底を評価する研究は著者の知る限り，行われていない．そこで，“実験 1-1” では，差異に基づくメッセージ特徴を用いて，評価実験を行う．

4.4.2 オンラインコミュニティの閲覧数

“実験 1-1” では，閲覧数をオンラインコミュニティの話題性を評価する指標に用いた．“実験 1-1” では，2 種類のオンラインコミュニティを評価する．まず，オンラインコミュニティ1 の閲覧数を図 4.6 および図 4.7 に示す．図 4.6 は，質問記事が投稿されてからの経過日数の散布図，図 4.7 は，閲覧数に対する返信数の散布図である．

図 4.6 および図 4.7 から，“実験 1-1” で評価するオンラインコミュニティ1 では，質問記事の閲覧数は正規分布の傾向が確認できた．図 4.6 の閲覧数と経過日数では，集中して閲覧数が多い時期がある．また，日数経過で閲覧数が増加する傾向が確認できた．相関係数は，0.28 であった．一方で，図 4.7 の閲覧数と返信数では，相関係数は，0.42 を示した．

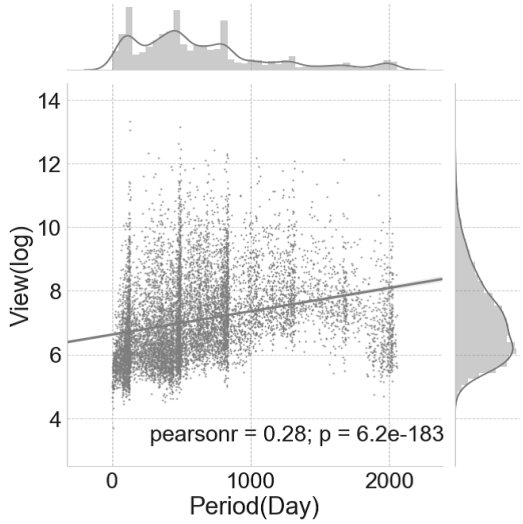

図 4.6: オンラインコミュニティ1 閲覧数と経過日数 (1)

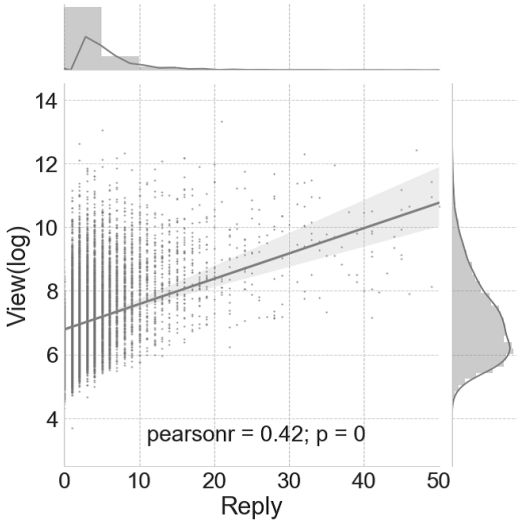

図 4.7: オンラインコミュニティ1 閲覧数と返信数 (1)

次に，オンラインコミュニティ2 の閲覧数を図 4.8 および図 4.9 に示す．図 4.8 は，質問記事が投稿されてからの経過日数の散布図，図 4.9 は，閲覧数に対する返信数の散布図である．

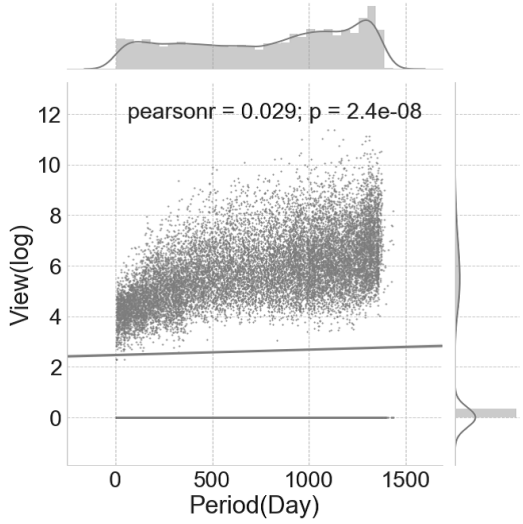

図 4.8: オンラインコミュニティ2 閲覧数と経過日数 (2)

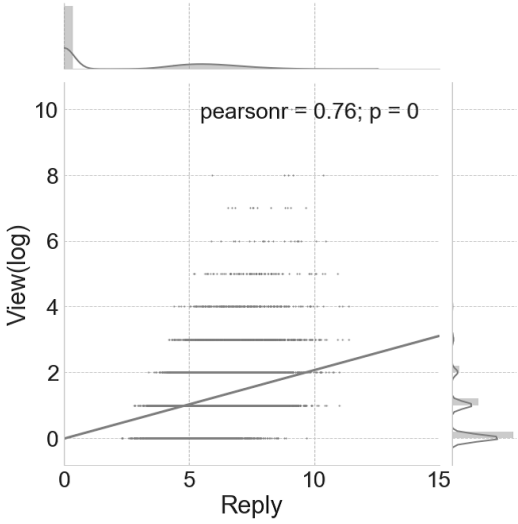

図 4.9: オンラインコミュニティ2 閲覧数と返信数 (2)

図 4.8 では，オンラインコミュニティ1 と異なり，時期に依存した閲覧数の増加はない．また，日数経過で閲覧数が増加する質問記事がある一方で，閲覧数が増加しない質問記事が多いことも明らかになった．閲覧数と経過日数の相関係数は，0.029 であった．図 4.9 では，閲覧数と返信数の相関係数は，0.76 という高い相関が明らかになった．

結果，2 種類のオンラインコミュニティで閲覧数と返信数には相関があり，コミュニケーションを増加させる要素であることが明らかになった．したがって，話題性の評価要素に閲覧数を用いることは妥当であると言える．

4.4.3 評価方法

“実験 1-1” ではメッセージ特徴の評価に 3.4.1 節の非線形回帰手法である SVR を用いた閲覧数の予測および 3.3 節の分類器，Ada Boost，Random Forests，MLP，K-NN を用いる．閲覧数の予測では，選択された基底における特徴量を用いて，特徴量選択の妥当性を 3.4.2 節の MAE，RMSE を算出し，評価する．分類器の評価では，閲覧数が多い質問記事と閲覧数が少ない質問記事の分類を 3.3.5 節の適合率，再現率，F-measure を算出し，評価する．分類器の評価で用いる特徴量は，4.3 節の差異に基づくメッセージ特徴を用いる．メッセージ特徴は，差異に基づいて評価した結果の基底である．非線形回帰手法や分類器の評価において，基底に寄与率の高い少数の特徴量で，質問記事の分類が行えた場合，話題性の評価に優れた特徴量である．したがって，特徴選択が妥当であるとし，基底の寄与率および抽出したメッセージ特徴が有用であるとした．

4.4.4 オンラインコミュニティ1 に対する基底選択と非線形回帰による評価

オンラインコミュニティ1 に対して，NMF の基底数を $M = 50$ に設定し，提案手法の適用を行った．提案手法による基底の評価結果を図 4.10 に示す．

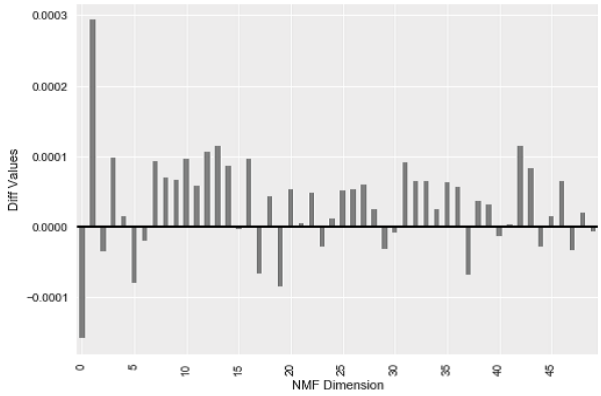

図 4.10: オンラインコミュニティ1 閲覧数に寄与率が高い基底

図 4.10 の横軸が基底 m の特性を表す集合 S_m の要素である．また，縦軸が集合間における係数の平均値の差である．縦軸の値が非負値の基底が，閲覧数が平均値以上のベクトル集合である L_1 の特性を表す基底である．結果，基底 1 が非負値で最も差異が大きいと評価された．したがって，基底 1 が閲覧数が平均値以上の集合 L_1 の特性を表す基底である．

得られた結果の基底 1 を基底選択して，SVR を用いて回帰分析を行い，閲覧数の予測を行い，予測誤差である MAE を算出した結果を図 4.12 および図 4.11 に示す．

また同様に，RMSE を算出した結果を図 4.14 および図 4.13 に示す．図の Input Feature は，回帰分析で用いる際の特徴量の数である．回帰分析の特徴量は，基底 1 に対する寄与率の順に用いている．

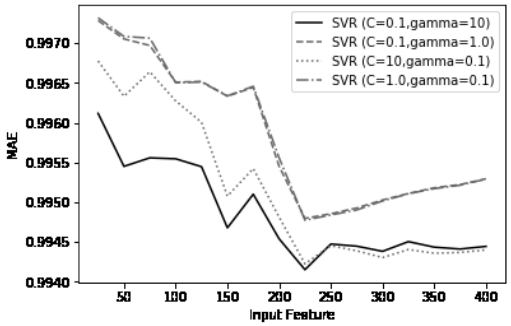

図 4.11: オンラインコミュニティ 1 回帰分析の結果を用いた MAE(1)

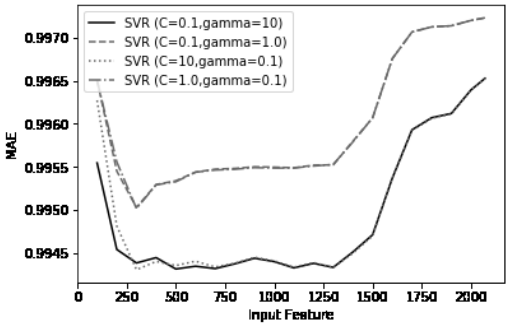

図 4.12: オンラインコミュニティ 1 回帰分析の結果を用いた MAE(2)

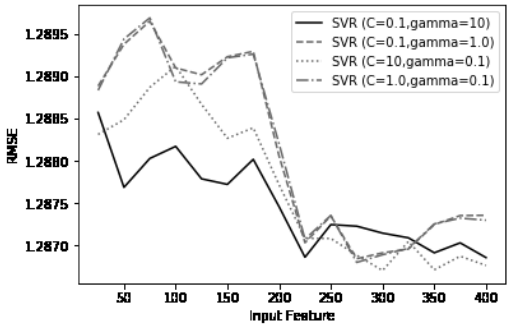

図 4.13: オンラインコミュニティ 1 特徴選択数と回帰分析の RMSE(1)

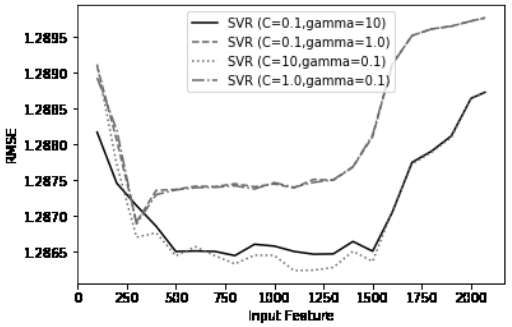

図 4.14: オンラインコミュニティ 1 特徴選択数と回帰分析の RMSE(2)

SVR の入力特徴選択数が 250 個前後が最も MAE の精度が良い (図 4.11)．一方で，1250-1500 個以上の入力特徴量がある場合，MAE の精度が下がる (図 4.12)．結果，すべての特徴量を用いた場合よりも，提案手法で得られた基底 1 に寄与率の高い少数の特徴量を特徴選択した場合の方が，精度が高い．

SVR の入力特徴選択数が 200-250 個が最も RMSE の値が最も良い (図 4.13)．一方で，1500 個以上の入力特徴量がある場合，RMSE の精度が下がる (図 4.14)．結果，MAE と同様に，すべての特徴量を用いた場合よりも，提案手法で得られた基底 1 に寄与率の高い少数の特徴量を特徴選択した場合の方が，精度が高い．したがって，提案手法の基底選択と基底に基づく寄与率による特徴量の選択は妥当であり，有用であることがデータで示された．

最後に，基底 1 に対する特徴量の寄与率上位の 10 個を表 4.4 に示す．

表 4.4: 基底 1 の寄与率上位 10 個の特徴量

	特徴量名		特徴量名
1	固有 (一般)	6	機能 (I-目的)
2	機能 (I-順接仮定)	7	ひらがな (頻度)
3	品詞 (助詞)	8	機能 (I-不可能)
4	意味 (2)(4.112)	9	機能 (I-依頼)
5	機能 (I-様態)	10	機能 (B-自然発生)

表 4.4 の「固有 (一般) 」は「固有名詞」,「機能 (*)」は「機能表現タグ」,「意味」は「意味分類コード」である．結果，固有名詞やひらがなの頻度の寄与が高い基底であることが明らかになった．意味分類コード (2)(4.112) の単語は,「その他の類 (展開)」の「ですから」「まして」「だから」などである．また，寄与率が高い特徴量に機能表現タグが複数含まれており，重要な特徴量であることが明らかになった．

4.4.5 オンラインコミュニティ1の文書分類

オンラインコミュニティ1より得られた特徴量を用いて文書分類実験を行った結果を表4.5から表4.8に示す.

表4.5: オンラインコミュニティ1 全特徴量

分類器	適合率	再現率	F値
AdaBoost	0.45	0.29	0.35
RandomForests	0.44	0.26	0.33
MLP	0.00	0.00	0.00
K-NN	0.42	0.37	0.39

表4.6: オンラインコミュニティ1 特徴量選択 (提案手法 基底の寄与率)

分類器	適合率	再現率	F値
AdaBoost	0.43	0.23	0.30
RandomForests	0.45	0.29	0.35
MLP	0.40	0.04	0.08
K-NN	0.44	0.42	0.43

表4.7: オンラインコミュニティ1 特徴量選択 (単変量特徴量選択)

分類器	適合率	再現率	F値
AdaBoost	0.46	0.28	0.35
RandomForests	0.45	0.28	0.34
MLP	0.00	0.00	0.00
K-NN	0.42	0.38	0.40

表 4.8: オンラインコミュニティ1 特徴量選択 (再帰的特徴量削減)

分類器	適合率	再現率	F 値
AdaBoost	0.44	0.29	0.35
RandomForests	0.45	0.27	0.34
MLP	0.00	0.00	0.00
K-NN	0.43	0.38	0.40

表 4.5 は，得られたすべての特徴量を用いて，文書分類した結果である．表 4.6 が，提案手法で得られた基底 1 の寄与率上位 100 個の特徴量を特徴選択し，文書分類した結果である．表 4.7 は，既存手法である単変量特徴量選択を用いて特徴量選択した結果であり，表 4.8 は，既存手法である再帰的特徴量削減を用いて特徴量選択し，文書分類した結果である．結果，すべての特徴量およびいずれの特徴量選択を用いた場合でも，オンラインコミュニティ1 の質問記事では，特徴量選択手法に関わらず，文書分類の精度は十分ではない．したがって提案手法ではなく，このため，オンラインコミュニティ1 においては，閲覧数の平均値を基に行った文書分類のラベリングの評価基準が妥当ではなかったと判断できる．

4.4.6 オンラインコミュニティ2 に対する基底選択と非線形回帰による評価

評価では，NMF の基底数を $M = 50$ に設定し，提案手法の評価を行った．提案手法による基底選択の結果を図 4.24 に示す．

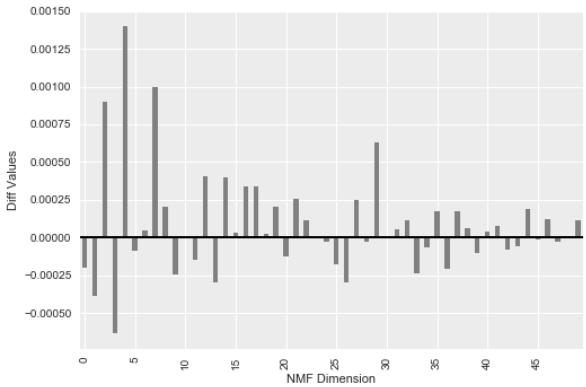

図 4.15: オンラインコミュニティ2 における閲覧数に寄与率が高い特徴量

NMF の基底数を M = 50 に設定し，提案手法の評価を行った．横軸が基底 m の特性を表す集合 Sm の要素である．また，縦軸が集合間における係数の平均値の差である．縦軸の値が非負値の基底が，閲覧数が平均値以上のベクトル集合である L1 の特性を表す基底である．追加実験では，基底 4 が非負値で最も差異が大きいと評価された．

得られた結果の基底 4 を用いて，SVR を用いて回帰分析を行い，閲覧数の予測を行い，予測誤差である MAE を算出した結果を図 4.17 および図 4.16 に示す．また同様に，RMSE を算出した結果を図 4.19 および図 4.18 に示す．図の Input Feature は，回帰分析で用いる際の特徴量の数である．また，回帰分析の特徴量は，基底 4 に対する寄与率の順に用いている．

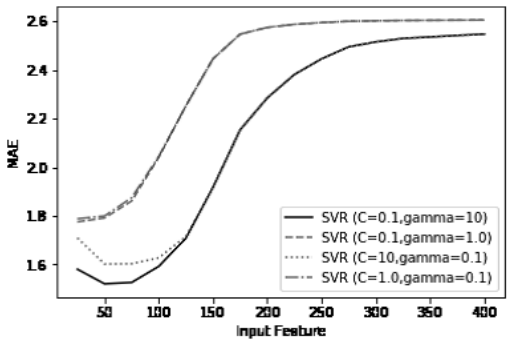

図 4.16: オンラインコミュニティ 2 回帰分析の結果を用いた MAE(1)

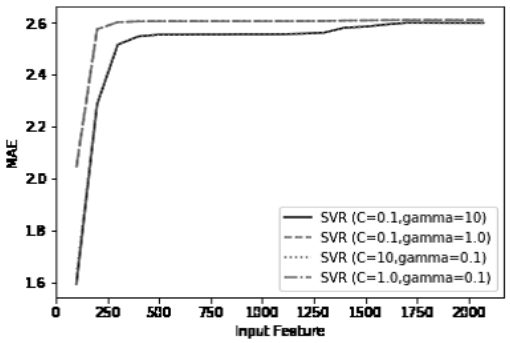

図 4.17: オンラインコミュニティ 2 回帰分析の結果を用いた MAE(2)

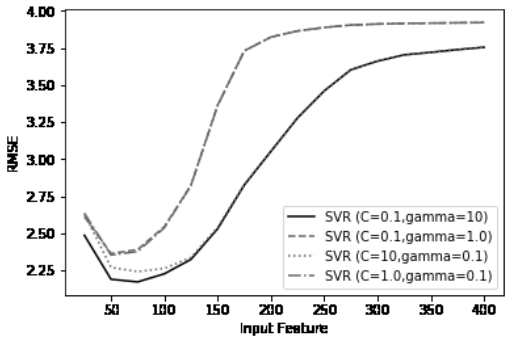

図 4.18: オンラインコミュニティ 2 特徴選択数と回帰分析の RMSE(1)

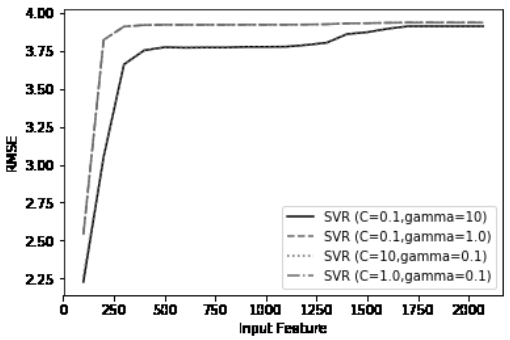

図 4.19: オンラインコミュニティ 2 特徴選択数と回帰分析の RMSE(2)

SVR の入力特徴選択数が 50-100 個前後が最も MAE の精度が良い (図 4.16)．一方で，100 個以上の入力特徴量がある場合，MAE の精度が下がる (図 4.17)．入力特徴選択数が 50 個である場合と 250 個以上である場合は，予測精度の差は明らかである．結果，すべての特徴量を用いた場合よりも，提案手法で得られた基底 4 に寄与率の高い少数の特徴量を特徴選択した場合の方が，精度が高い．同様に，SVR の入力特徴選択数が 50-100 個が RMSE の精度が良い (図 4.18)．MAE と同様，に入力特徴選択数が 50 個である場合と 250 個以上である場合は，予測精度の差は明らかである (図 4.19)．

結果，すべての特徴量を用いた場合よりも，提案手法で得られた基底 4 に寄与率の高い少数の特徴量を特徴選択した場合の方が，回帰分析の予測精度が高い．したがって，提案手法の基底選択と基底に基づく寄与率による特徴量の選択は妥当であり，有用であることがデータで示された．

最後に，基底 4 に対する特徴量の寄与率上位の 10 個を表 4.9 に示す．

表 4.9: 基底 4 の寄与率上位 10 個の特徴量

	特徴量名		辞書の単語
1	TOPIC(LDA)	6	意味 (2)(1.353)
2	意味 (2)(2.304)	7	意味 (2)(4.35)
3	意味 (2)(1.304)	8	意味 (2)(1.564)
4	意味 (2)(1.366)	9	意味 (2)(3.133)
5	意味 (2)(4.314)	10	TOPIC(LDA)

表 4.9 の「意味」は「意味分類コード」である．結果，意味分類コードの寄与が高い基底であることが明らかになった．

4.4.7 オンラインコミュニティ2 の文書分類

オンラインコミュニティ2 より得られた特徴量を用いて文書分類実験を行った結果を表 4.10 から表 4.13 に示す．

表 4.10: オンラインコミュニティ2 全特徴量

分類器	適合率	再現率	F 値
AdaBoost	0.89	0.87	0.88
RandomForests	0.87	0.78	0.82
MLP	0.92	0.88	0.90
K-NN	0.83	0.75	0.78

表 4.11: オンラインコミュニティ2 特徴量選択 (提案手法 基底の寄与率)

分類器	適合率	再現率	F 値
AdaBoost	0.87	0.83	0.85
RandomForests	0.87	0.82	0.84
MLP	0.87	0.87	0.87
K-NN	0.85	0.80	0.83

表 4.12: オンラインコミュニティ2 特徴量選択 (単変量特徴量選択)

分類器	適合率	再現率	F 値
AdaBoost	0.88	0.85	0.87
RandomForests	0.88	0.83	0.85
MLP	0.00	0.00	0.00
K-NN	0.87	0.83	0.85

表 4.13: オンラインコミュニティ2 特徴量選択 (再帰的特徴量削減)

分類器	適合率	再現率	F 値
AdaBoost	0.89	0.86	0.87
RandomForests	0.88	0.84	0.86
MLP	0.00	0.00	0.00
K-NN	0.88	0.82	0.85

表 4.10 は，得られたすべての特徴量を用いて，文書分類した結果である．表 4.11 が，提案手法で得られた基底 1 の寄与率上位 100 個の特徴量を特徴選択し，文書分類した結果である．表 4.12 は，既存手法である単変量特徴量選択を用いて特徴量選択した結果であり，表 4.13 は，既存手法である再帰的特徴量削減を用いて特徴量選択し，文書分類した結果である．結果，オンラインコミュニティ2 では，正解クラスへの分類結果は 0.8 以上の高い分類精度が得られた．したがって，文書分類に有用であることが明らかである．このため，オンラインコミュニティ 2 においては，閲覧数の平均値を基に行った文書分類のラベリングの評価基準が妥当であると判断できる．

4.4.8 話題性の予測に関する考察

評価に用いたオンラインコミュニティの考察から，始める．まず，オンラインコミュニティ 1 は，Apple サポートコミュニティのデータセットである．Apple サポートコミュニティは，コンシューマ[21]が，直接質問記事を投稿する．また，利用者層の幅は非常に広い．したがって，メディアとして考察した場合は，ゼネラル・メディアに相当する性質を持つ．一方で，オンラインコミュニティ2 は，stackover flow のデータセットである．stack overflow は Stack Exchange の Q&A コミュニティの一つであり開発者向けのコミュニティである．stack overflow は専門家や熟練者など，高度な知識を持つ，エキスパートが利用者である．このため，メディアとして考察した場合は，クラス・メディアに相当する性質を持つ．

[21]商品やサービスの最終的な利用者，消費者のこと．消費者という概念はある商品やサービスを直接利用する人であるが，購買行為を決定する人も含めて消費者と呼ぶこともある [2].

4.4.4 節の結果から，各オンラインコミュニティで，閲覧数と経過日数の相関係数が大きく異なることが明らかになった．また，閲覧数と返信数の相関係数においても，オンラインコミュニティ1 で 0.42，オンラインコミュニティ2 で 0.76 である．ゆえに，質問記事を閲覧した際の返信度合いは，オンラインコミュニティのメディアの性質で異なると推定される．ここで，オンラインコミュニティの一つである不満買取センター[22] の不満カテゴリ辞書データを用いて，オンラインコミュニティのコンテンツを考察する．不満カテゴリ辞書は，2015 年 3 月 18 日から，2016 年 12 月 1 日までのノイズを排除した投稿記事 3,527,336 件から作成されている [132]．辞書のエントリ数は 953,776 件であり，複数カテゴリに登録されている重複エントリを除いたエントリ数は 110,866 件である．不満買取センターに不満投稿が投稿されるカテゴリと，特徴的な単語の例を表 4.14 に示す．単語は，不満カテゴリ辞書の TF-IDF [23] におけるスコア上位エントリの名詞である．

表 4.14: 不満買取センターのカテゴリと単語の例 [132]

カテゴリ名	単語の例
暮らし・住まい	布団，雨，枕
ファッション	腕時計，靴
趣味・エンタメ	映画，CD
食品・飲料	グミ，ワイン
外食・店舗	弁当，居酒屋
医療・福祉	整体，介護，薬
アウトドア・スポーツ	バイク，自転車
デジタル・家電	録画，デジカメ
宿泊・観光・レジャー	ホテル，部屋
公共・環境	バス，飛行機
教育	幼稚園，保育園
国際・文化	留学，日本
政治・行政	選挙，政治家
人間関係	離婚，結婚，人
仕事	転職，仕事，面接
ペット	ペットショップ

[22]不満買取センター : https://fumankaitori.com

[23]単語の重み付け技法の一つ．文書内単語の相対的な重要性を表す正規化されたスコア．単語の頻度と文書頻度の逆数で算出.

ところで，不満カテゴリ辞書データの特徴的な単語は，登録カテゴリで異なる．また辞書では，飲食物の商品名など，商品の固有名詞も登録されており，商品名は食品・飲料や，宿泊・観光・レジャーなど，関連性のある複数カテゴリで登録されている．一方で，他のカテゴリで登録されている特徴的な単語であっても，関連性がないカテゴリでは登録されていない．ゆえに，同一のオンラインコミュニティであっても，コンテンツが広範なオンラインコミュニティの場合，オンラインコミュニティの各カテゴリが，クラス・メディアに相当する場合もあると推定される．評価実験の話題性の予測では，オンラインコミュニティ1 およびオンラインコミュニティ2 で，特徴量選択は予測誤差の精度は有用な結果であった．しかし，オンラインコミュニティ2 の予測精度の向上は明らかである一方で，オンラインコミュニティ1 の予測精度の向上は限定的であった．4.4.4 節の結果から，オンラインコミュニティ1 は，ゼネラル・メディアに相当すると推定される．ゼネラル・メディアの性質を持つオンラインコミュニティの場合，コンテンツの広範であり，カテゴリが多い．したがって，表 4.14 のように，カテゴリが異なった場合，特徴的な単語は異なると推定される．

“実験 1-1” で用いたトピックモデルアルゴリズムは，カテゴリ分類においても有用なアルゴリズムである．提案手法の適用の結果，オンラインコミュニティ1 では，閲覧数予測においては，有用性は限定的であった．ゼネラル・メディアに近い性質を持つオンラインコミュニティ1 は，利用者が広範である．したがって，閲覧数が増加しやすい傾向や特徴的な単語が不明瞭であることなどが推定される．このため，オンラインコミュニティ1 の閲覧数予測の精度向上には，カテゴリ分類後に，閲覧数予測を行う必要があると推定される．

一方で，オンラインコミュニティ2 は 4.4.4 節の結果から，クラス・メディアの性質を持つと推定される．オンラインコミュニティ2 はオンラインコミュニティ1 と比較して，利用者が限られた範囲である．提案手法の適用の結果，オンラインコミュニティ2 では，閲覧数予測において，高い有用性が明らかになった．これは，クラス・メディアの性質を持つオンラインコミュニティ2 は，利用者が限られており，閲覧数が増加しやすい一定の傾向が存在することや，あるいは閲覧数が増加する特徴的な単語が明瞭であることなどが推定される．加えて，コンテンツも限られた範囲であり，カテゴリが異なった場合でも，特徴的な単語は類似の傾向があると推定される．

ゆえに，SVR で回帰分析を行った際に，クラス・メディアであるオンラインコミュニティ2 は明瞭な結果となり，ゼネラル・メディアであるオンラインコミュニティ1 では有効ではあるが限定的な結果であったと推定される．

4.4.9 話題性の判別に関する考察

分類実験では，ゼネラル・メディアに相当するオンラインコミュニティ1 とクラス・メディアに相当するオンラインコミュニティ2 で結果が大きく異なった．まず，既存手法で抽出した特徴量を用いて，文書分類を行った結果を考察する．

“実験 1-1” の文書分類では，質問記事の分類の基準となるラベルは，閲覧数の平均値を基にラベリングした．オンラインコミュニティ1 の文書分類では，すべての特徴量を用いた場合や，既存研究の特徴量選択手法を用いた場合など，いずれの手法を用いた場合でも，正解クラスへの分類結果は F 値で，0.3 から 0.4 程度である．パラメータの最適化を行わない場合，MLP の分類器では分類困難である．

次に，オンラインコミュニティ1 の層化 5 分割交差検証の交差検証スコアを表 4.15 に示す．表 4.15 の RF は，RandamForests である．

表 4.15: 交差検証 (オンラインコミュニティ1)

分類器	1	2	3	4	5
AdaBoost	0.54	0.53	0.52	0.54	0.53
RF	0.52	0.54	0.53	0.52	0.52
MLP	0.55	0.55	0.56	0.54	0.55
K-NN	0.50	0.52	0.52	0.51	0.51

表 4.15 より，交差検証スコアの場合でも，結果は十分ではない．したがって，オンラインコミュニティ1 では，基底選択に関わらず，十分な文書分類が行えていないと言える．このため，オンラインコミュニティ1 では，特徴量選択ではなく，正解クラスのラベリング基準を最適値にする必要がある．

一方で，オンラインコミュニティ2 では，既存研究の特徴量選択手法を用いた場合など，いずれの手法を用いた場合でも，正解クラスへの分類結果は F 値で，0.7 から 0.9 の範囲である．

次に，オンラインコミュニティ1 の層化 5 分割交差検証の交差検証スコアを表 4.16 に示す．表 4.16 の RF は，RandamForests である．

表 4.16: 交差検証 (オンラインコミュニティ2)

分類器	1	2	3	4	5
AdaBoost	0.86	0.86	0.87	0.87	0.87
RF	0.86	0.85	0.87	0.86	0.87
MLP	0.89	0.88	0.89	0.89	0.90
K-NN	0.84	0.84	0.85	0.85	0.85

表 4.15 より，交差検証スコアの場合でも，有用な結果であった．したがって，オンラインコミュニティ2 では，閲覧数のラベリング基準は平均値が妥当であったことが明らかになった．

ここで，提案手法は，目標変数である閲覧数から，有用な基底を明らかにし，基底選択を行う手法である．文書分類で有用な結果が得られたオンラインコミュニティ2 の結果を用いて，提案手法による基底選択の有用性を考察する．

基底選択の結果，基底の寄与率上位の特徴量で，重要な特徴量が明らかになり，特徴量の評価が行える．したがって，重要な特徴量のみを用いる特徴量選択が行える．分類結果で考察した場合，既存手法である単変量特徴量選択および再帰的特徴量削減では，分類困難であった MLP による分類が行えた．したがって，分類器に依存せずに，少ない特徴量で，すべての特徴量を用いた場合に相当する結果が得られた．このため，分類器に依存せず有用な特徴量選択が行える手法である．また，分類精度だが，提案手法を用いたと比較し，すべての特徴量を用いた場合や，既存手法の特徴量選択を行った場合の方が，適合率や再現率，F 値の結果は良い．しかし一方で，値の差は誤差の範囲であり，すべての特徴量を用いた場合と同程度の分類結果であり，特徴量選択手法としての性能を有している．加えて，既存の特徴量選択手法と，同程度の結果も得られた．

特徴量選択としては，基底に対する特徴量の寄与率で，統計的な関係を基にした単変量特徴量選択と同様に，関係がないノイズに相当する特徴量の除去が行える．また，計算量基準では，再帰的特徴量削減は，特徴量の次元数を指定する場合，特定の次元数になるまで，繰り返しモデルベースを適用する．したがって，非常に大きな計算量を必要とする．一方で，提案手法は基底を選択した場合に，基底のベクトルである寄与率で特徴量を評価している．したがって，計算量は，再帰的特徴量削減と比較して少ない．このため，計算量基準では，既存手法である再帰的特徴量削減と比較し，有用な特徴量選択である．ゆえに，文書分類においては，提案手法は分類器に依存せず，再帰的特徴量削減と比較して，少ない計算量で特徴量選択が行える手法であると言える．

文書分類で有用な結果が得られたオンラインコミュニティ2で，提案手法で得られた基底の特徴量を用いた結果の Precision-recall カーブ，受信者動作特性 (ROC)，AUC の結果を図 4.20，図 4.21 に示す．AUC は，ROC のカーブ下の領域である．

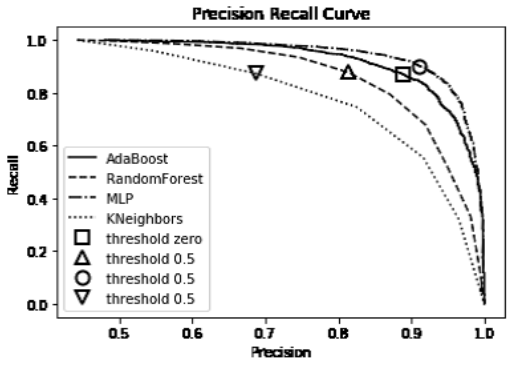

図 4.20: 実験 1-1 オンラインコミュニティ 2 Precision-Recall カーブ

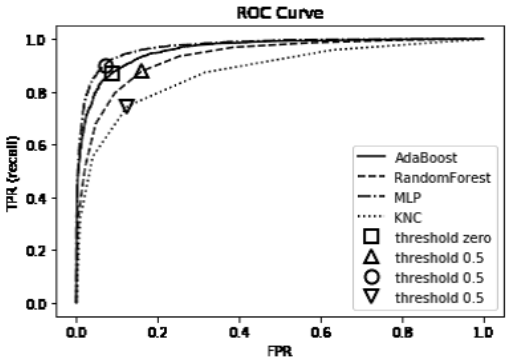

図 4.21: 実験 1-1 受信者動作特性 (ROC)

表 4.17: 実験 1-1 における AUC

Classification methods	AUC
AdaBoost	0.9608
RandomForests	0.9313
MLP	0.9725
K-NN	0.8747

結果，分類スレッショルドの最適化を行うことで，適合率や再現率，F 値などの最適化が行える特徴量選択であることが明らかになった．したがって，提案手法による基底選択は，文書分類に有用な特徴量を評価する手法としての有用性があると言える．また，結果から，AUC では，MLP が最も良いスコアの分類器であることが明らかになった．

4.5 - 実験 1-2 -
伝播メディアのコミュニケーションにおける平易化の課題への応用

4.5.1 概要

“実験 1-2” では，異なるテキストコーパス[24] と既存研究の特徴量で平易化テキストに関連する差異にメッセージ特徴を明らかにする．近年，ICT 技術の発展で誰でもが情報を発信・共有できる基盤が整い，多様なユーザが情報発信を行っている．結果，インターネットでは同一の内容でも難易が混在した多様なコンテンツが存在している．したがって，同一の内容でも自分が理解できる文書を探し当てるのには，相当な時間と労力を要する．また，トップダウン形式[25]．の情報発信の場合，理解できない受信者が多い場合，影響は限定的である．このため，情報発信において，受信者が理解できるメッセージ特徴が重要である．

“実験 1-2” では，同一の発行者が異なる対象読者に向けて発行した報告書の違いを明らかにすることに取り組む．ここで，特別なスキルを求めない幅広い読者層に向けて発行した普及啓発書を「平易化文書」と称する．そして専門的な知識を有する実務者向けの報告書を比較に用いる．平易化文書に関連した特徴量が明らかになることで，コンテンツの難易を自動で判別し，特別なスキルを求めない幅広い読者層に向けてより良い情報発信が期待できる．

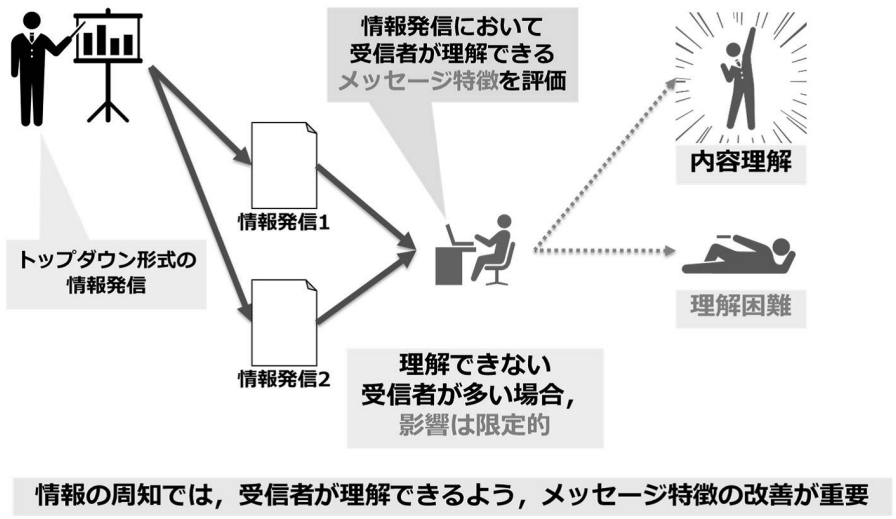

図 4.22: 研究背景 (実験 1-2)

[24] コーパスは言語資料体．個別言語・発話などの情報を大規模または網羅的に集めたもの．電子化された例文データベースの場合，例文中に語句の品詞，例文，発音などの情報が付加される．

[25] マクロな特性が与えられ，特性を生成するミクロな要素の関係をつけ，作られる情報．組織の上位から下位へ命令が伝達される管理方式など．対して，ボトムアップ型は多数の要素が相互作用を通じて，マクロ的な秩序現象が創発し，作られる情報．

“実験 1-2” のメッセージ特徴抽出手順の概要を図 4.23 に示す．図 4.23 の (1) から (4) が，“実験 1-2” のメッセージ特徴抽出手順を示すステップである．

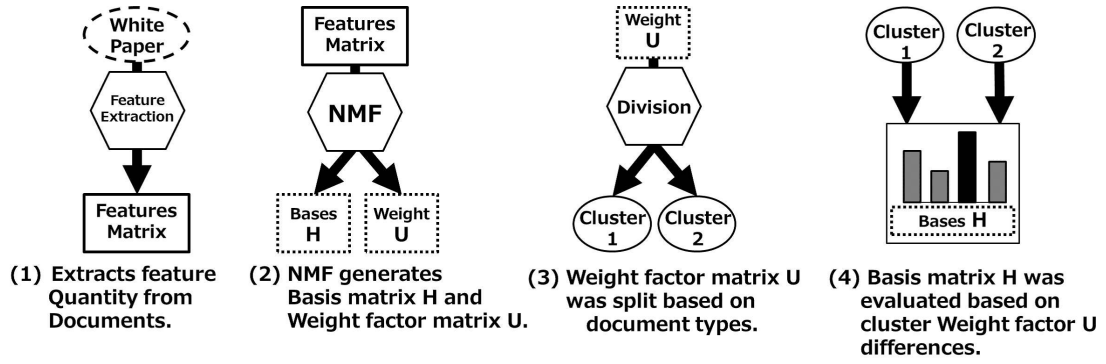

図 4.23: 実験 1-2 メッセージ特徴抽出手法

図 4.23(1) の特徴量の抽出では 4.2.1 節，4.2.2 節，4.2.3 節に示した特徴量を抽出している．抽出する特徴量は既存研究の特徴量を組み合わせた，31 個，各特徴量を水平方向に連結した際の次元数は 2,071 次元である．そして (2) から (4) が，“実験 1-2” のを評価する手順である．(2) では NMF を用いて行列分解を行い，特徴量の変換を行う．(2) および (4) のメッセージ特徴抽出の基底選択では，4.3 節の差異による基底選択を用いた．

“実験 1-2” は，2.2 節で示した従来研究と比較し，テキスト情報の特徴量を網羅的に抽出している．また，テキスト平易化の従来研究で，NMF による行列分解で特徴量変換し，網羅的なテキスト情報から基底を評価する研究は著者の知る限り，行われていない．そこで，“実験 1-2” では，差異に基づくメッセージ特徴を用いて，評価実験を行う．

4.5.2 評価対象

“実験 1-2” では，環境省が発行している年次報告書である環境白書を用いる．環境白書は，環境状況の報告と環境保全に関する施策で構成されている．最新の環境白書は，循環型社会白書や生物多様性白書と合本し，環境・循環型社会・生物多様性白書[26] となっている．また，環境白書の普及啓発冊子として，英語版の環境白書や図で見る環境白書，こども環境白書[27]．も発行されている．こども環境白書は，当該年度の環境白書の代表的な環境問題から，わかりやすいトピックを抽出した年次報告書である．敷衍を目的に，平易な表現で解説している普及啓発冊子であることから，本研究では，こども環境白書を平易化コーパスと称する．こども環境白書では,「地球がどんどんあたたまる (2016 年 P.3)」など理解しやすい表現が用いられている．“実験 1-2” における評価対象は，環境白書およびその普及啓発冊子であるこども環境白書を対象する．

[26]環境省 環境白書・循環型社会白書・生物多様性白書 : https://www.env.go.jp/policy/hakusyo/
[27]環境省 こども環境白書 : https://www.env.go.jp/policy/hakusyo/kodomo.html

年次報告書は 2009 年から 2016 年までの 8 年分が評価対象である．年次報告書は PDF で提供されているため，PDF よりテキスト情報を抽出した[28]．また，一部こども環境白書 (2009 年，2011 年，2014 年) に関しては，手作業および WebService を使用してテキスト情報を抽出している[29]．

年次報告書は「。」「？」「！」で改行し，文書とした．テキスト情報の前処理では，文字コードを UTF-8 に変換し，改行文字とバイト列の欠損行は削除した．8 年分の環境・循環型社会・生物多様性白書とこども環境白書の合計文書数は 51,689 件である．年次報告書より抽出する特徴量の次元数は 2,071 次元である．したがって，“実験 1-2” における NMF の観測行列 Y は，51,689 行 2,071 列の長方行列である．ここで，環境・循環型社会・生物多様性白書は，難解なコーパスに相当する．また，こども環境白書は，平易化コーパスに相当する．“実験 1-2” のメッセージ特徴の抽出では，4.3 節の差異による基底選択を用いている．したがって，“実験 1-2” ではコーパスの難易度の差異で平易化コーパスに影響がある，メッセージ特徴を評価する．

4.5.3 評価方法

“実験 1-2” ではメッセージ特徴の評価に 3.3 節の分類器，Ada Boost，Random Forests，MLP，K-NN を用いる．分類器の評価では，難解なコーパスと平易化コーパスの分類を 3.3.5 節の適合率，再現率，F-measure を算出し，評価する．分類器では，4.3 節の差異に基づいたメッセージ特徴を用いる．メッセージ特徴は，差異に基づいて評価した結果の基底である．得られた基底に寄与率の高い少数の特徴量で，コーパスの分類が行えた場合，コーパスの分類に有用な特徴量である．したがって，特徴選択が有用であるとし，基底の寄与率および抽出したメッセージ特徴が有用であるとした．“実験 1-2” では分類器による評価に加え，因果関係の推定を行う．因果関係の推定では，3.5 節のベイジアンネットワークを用いる．文書の種別に基づき，寄与率上位の特徴量を用いて，平易化コーパスの判断に因果関係がある特徴量を明らかにする．次に評価結果として，メッセージ特徴を 4.5.4 節に，分類器の評価結果を 4.5.5 節に，受信者動作特性 (ROC) と AUC の結果を 4.5.6 節に，ベイジアンネットワークによる因果関係の推定結果を 4.5.7 節に示す．

[28]PDF ファイルからテキストおよび画像を抽出する方法（Acrobat DC）：
https://helpx.adobe.com/jp/acrobat/kb/cq06200852.html

[29]PDF TXT 変換 PDF を Text に : https://pdftotext.com/ja/

4.5.4 基底選択の評価結果

評価では，提案手法の基底数を $M = 50$ に設定し，特徴量の評価を行った．提案手法による基底の評価結果を図 4.24 に示す．縦軸が集合間の係数値の差である．集合の係数値は平均値である．横軸が基底番号である．縦軸の値が非負値の基底が，平易化コーパスに対しての係数値が大きい基底である．

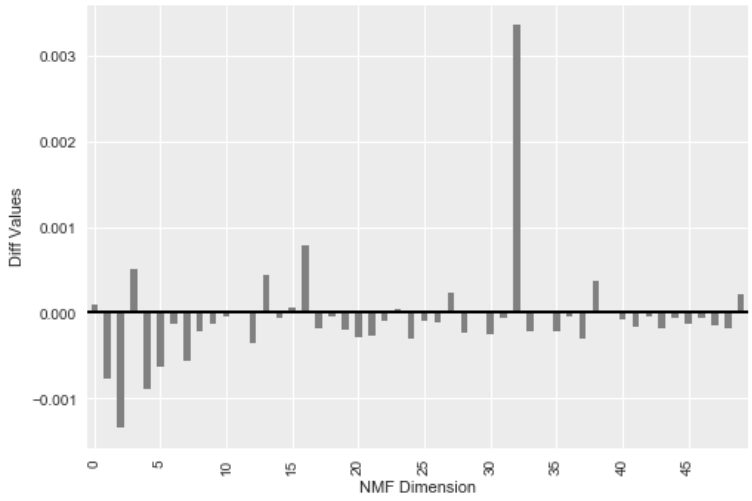

図 4.24: 平易化コーパスに寄与率が高い特徴量と状況変数

NMF の行列分解では，係数行列 U の成分がスパースになる傾向がある [112]．図 4.24 の結果から，基底 32 が非負値で最も差異が大きい．したがって，基底 32 は平易化コーパスへの係数の重みが大きく，もう一方の実務者向けの報告書への重みが小さい．結果，基底 32 は平易化コーパスの特性を表す基底である．得られた基底 32 の特徴量の上位 20 個を表 4.18 に示す．寄与率による評価では，ひらがな比率が最も寄与率が高く評価された．ひらがな比率は既存研究の平易化コーパスでも関連した特徴量であるため，評価された基底は平易化コーパスに関連した基底である解釈できる．また，機能表現や TOPIC(LDA) の特徴量が寄与率が高く評価された．加えて，拡張固有表現 (鳥類) など環境白書の内容に関連した特徴量も評価された．

表 4.18: 各分類手法の評価指標

基底 32 の特徴量の上位 20 個	
ひらがな比率	品詞比率 (副詞)
TOPIC(LDA)	機能表現 (I-推量-高確実性)
感情極性 (ポジティブ比率)	拡張固有表現 (鳥類)
TOPIC(LDA)	機能表現 (B-意志)
機能表現 (I-不可避)	TOPIC(LDA)
機能表現 (B-否定)	意味分類 (2)(1.565)
機能表現 (I-不許可)	機能表現 (I-理由)
機能表現 (I-推量-不確実)	機能表現 (B-勧誘)
機能表現 (I-勧誘)	機能表現 (I-判断)
機能表現 (I-疑問)	意味分類 (2)(1.164)

4.5.5 平易化コーパスの分類評価の結果

提案手法を適用し，得られた係数値が大きい基底 32 の寄与率上位 100 個の特徴量を用いて分類器による文書群の分類評価を実施した．分類器による分類結果を表 4.19 に示す．

表 4.19: 各分類手法の評価指標

分類手法	適合率	再現率	F-measure
AdaBoost	0.80	0.67	0.73
RandomForests	0.92	0.67	0.78
MLP	0.88	0.66	0.76
K-NN	0.90	0.50	0.64

結果，すべての分類器で「適合率」の高い指標値が得られた．最も「適合率」が高い分類器は「Random Forests」の「0.92」である．次いで「K-NN」の適合率が高い．一方で,「再現率」では「MLP」が最も良い．結果，適合率と再現率の調和平均値である「F-measure」では，「RandomForests」が最も高い「0.78」を示した．

平易化コーパスと実務者向けの報告書は，発行年度で年度ごとに内容が異なる．そこで，各年度に分けて，分類器による分類を行った結果を評価指標ごとに図 4.25，図 4.26，図 4.27 に示す．

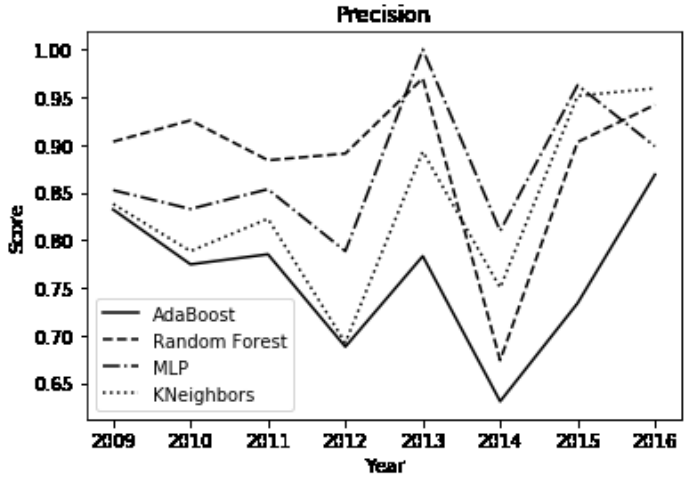

図 4.25: 各年度の Precision

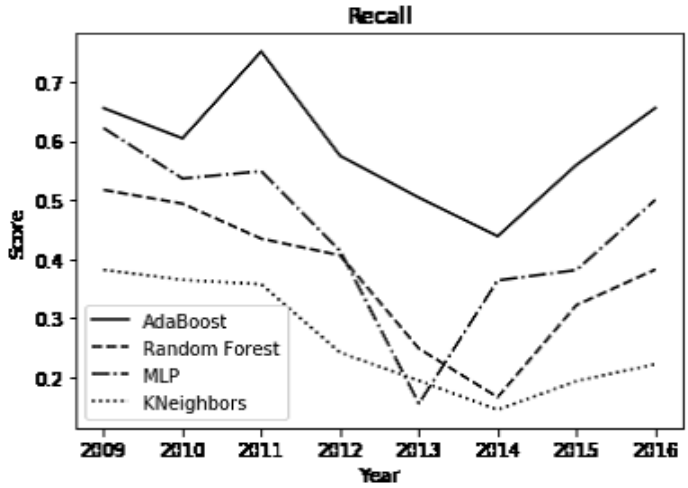

図 4.26: 各年度の Recall

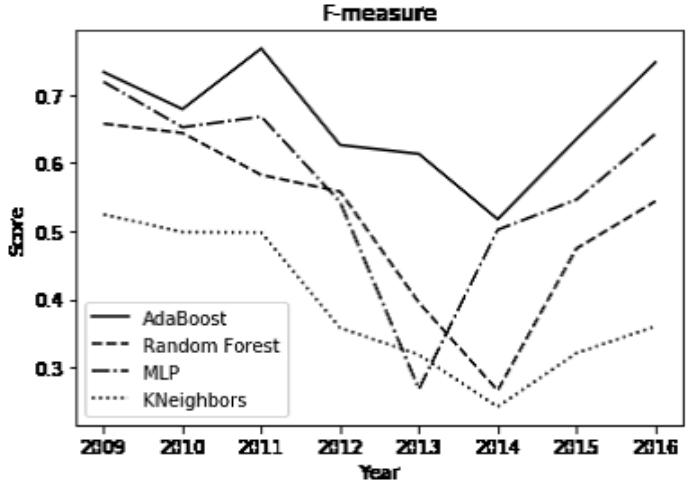

図 4.27: 各年度の F-measure

結果，各年度で，値は異なるが，すべての年度で Precision が高いが Recall は低い傾向が見られた．したがって，調和平均値である F-mesuare を算出した場合，Recall の影響が大きいことが明らかになった．分類手法別では,「Precision」の結果では,「RandomForests」の分類性能が高い．一方で「Recall」の結果では,「AdaBoost」の分類性能が高い．このため,「F-measure」では，すべての年度で「AdaBoost」の分類性能が高いという結果が得られた．このことから，分類評価はすべての年度のデータセットで評価した場合でも，各年度データセットで評価した場合でも有用性のある結果が得られた．したがって，提案手法の基底選択は有用性があると判断できる．

4.5.6 受信者動作特性 (ROC) と AUC

4.5.5 節で得られた結果のうち，各年度で比較した場合，Recall の影響が大きいことが明らかになった．Recall は分類器のパラメータである決定スレッショルドを変更し，平易化コーパスに分類される割合を増加変化させることで，再現率を向上させることができる．学習用のデータ (75%) でモデルを作成し，評価用のデータ (25%) で評価を行った場合の「適合率 - 再現率カーブ」を図 4.28,「受信者動作特性 (ROC)」を図 4.29，AUC を表 4.20 に示す．図 4.28 は縦軸が再現率，横軸が適合率である．また，図 4.29 は縦軸が再現率，横軸が偽陽性率である．

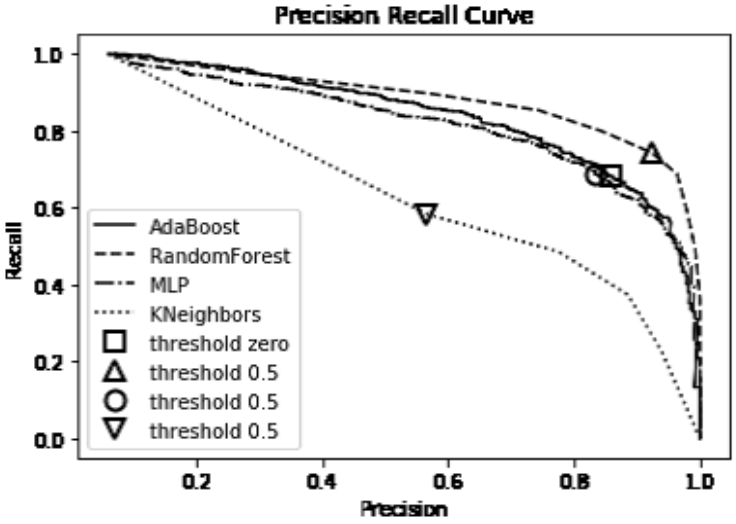

図 4.28: 実験 1-2 Precision - Recall カーブ

表 4.20: 実験 1-2 における AUC

分類手法	AUC
AdaBoost	0.9741
RandomForests	0.9624
MLP	0.9581
K-NN	0.8578

結果，図 4.28 および，図 4.29 から，分類器である「AdaBoost」や「RandomForests」,「MLP」などは，決定スレッショルドを最適化することで，より良い再現率が得られることが明らかになった．AUC は表 4.20 の受信者動作特性 (ROC) のカーブ下の領域の値である．AUC は分類器が機能しているかの指標である．結果，すべての分類器で良い値が得られたことが明らかになった．

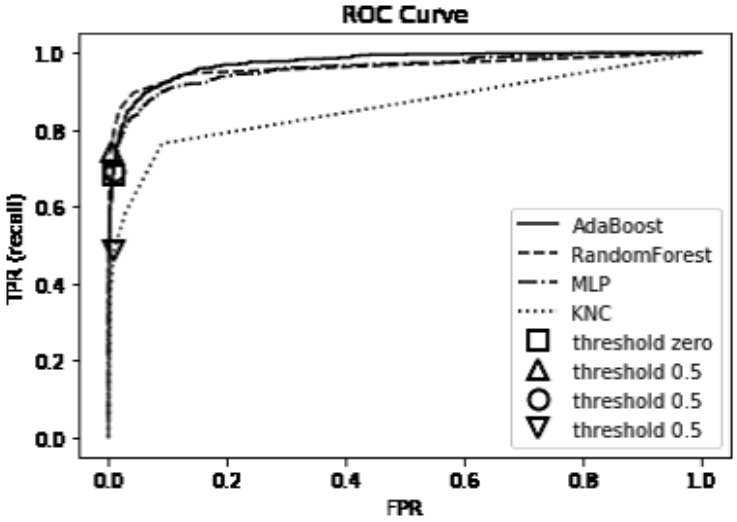

図 4.29: 実験 1-2 受信者動作特性 (ROC)

4.5.7 評価実験 (特徴量の推定)

文書群と特徴量との因果関係をベイジアンネットワークで評価した結果を表 4.21 に示す．分類器による評価と同様に，特徴量には提案手法を適用し，得られた係数値が大きい基底 32 の寄与率上位 100 個の特徴量を用いた．

表 4.21: 因果関係がある特徴量

No.	基底 32
1	機能表現タグ (B-否定)
2	機能表現タグ (I-不可能)
3	機能表現タグ (I-推量-不確実)
4	機能表現タグ (B-勧め)
5	読点間距離
6	TOPIC(LSI)
7	機能表現タグ (B-勧誘)
8	機能表現タグ (B-並立)

結果，ベイジアンネットワークの推定では，機能表現タグ，読点間距離，TOPIC(LSI) が因果関係のある特徴量として評価された．結果のうち，読点間距離に関しては，読みやすさの既存研究ので用いられている特徴量である．したがって，基底 32 の上位の特徴量には，平易化コーパスに関連性のある特徴量が含まれていることが明らかになった．

4.6 むすび

本章では，非負値行列因子分解アルゴリズムから得られるメッセージ特徴を，ベイトソンの情報の定義と組み合わせることで，伝播メディアにおける評価集合の差異における問題の改善が期待できるメッセージ特徴の選択手法を提案した．

そして，選択手法で得られたメッセージ特徴を伝播メディアにおける話題性および伝播メディアにおけるコミュニケーションの平易化で，性能評価実験を行った．伝播メディアにおける話題性の評価では，オンラインコミュニティの閲覧数に基づいた 2 種類の文書集合のメッセージ特徴を評価した．評価実験では，2 種類のオンラインコミュニティを評価した．また，伝播メディアにおけるコミュニケーションの平易化の評価では，平易化されたテキストを用いて，メッセージ特徴を評価した．

以下，本章で得られた結果を示す．

- 伝播メディアにおける話題性の課題に対する性能評価実験では，2,000 次元を越える高次元の特徴量から有効なメッセージ特徴を評価した．結果，メッセージ特徴に基づいた特徴量で，クラス・メディアのオンラインコミュニティにおいて優れた分類精度結果が得られた．また，非線形回帰においても，クラス・メディアのオンラインコミュニティにおいて，明瞭な予測精度の向上が得られた．したがって，提案手法は，伝播メディアにおける話題性の課題において，有効なメッセージ特徴の選択手法であると言える．

- 伝播メディアにおけるコミュニケーションの平易化の課題に対する性能評価実験においても，2,000 次元を越える高次元の特徴量から有効なメッセージ特徴を評価した．結果，メッセージ特徴に基づいた特徴量で，優れた分類精度が得られた．また，因果関係分析では，既存研究と同様の結果が得られた．したがって，提案手法は，伝播メディアにおけるコミュニケーションの平易化の課題において，有効なメッセージ特徴の選択手法であると言える．

- 伝播メディアにおける評価集合の差異における問題において，2 種類の評価実験で，優れた分類精度を示したことから，本研究におけるメッセージ特徴の有効性，また，選択手法においても優れた結果が得られたと言える．

- 話題性の課題に対する性能評価実験において，既存の特徴量選択手法と比較した場合においても，分類器依存せずに優れた分類精度が得られた．したがって，特徴量選択手法として比較した場合においても，優れた比較結果が得られたと言える．

ここで，差異に基づくメッセージ特徴の特徴選択手法では，2 種類の伝播メディアの集合を用いて，性能評価実験を行った．本研究においては，テキストコミュニケーションメッセージで性能評価実験を行った．評価実験の結果から異なるテキストコミュニケーションメッセージにおいても適用可能であり，伝播メディアをはじめとするテキストコミュニケーションにおいて，幅広い応用が期待できる．

第5章　回帰に基づくメッセージ特徴の特徴選択手法とその性能評価

5.1　はじめに

科学技術に基づいた場所に依存しないコミュニケーションの増加で，情報の取捨選択の機会の増加し，受信者の想定が困難となった．回帰に基づくメッセージ特徴では，利用者が有益と判断するコンテンツのメッセージ特徴を受信者の想定の基準となる目標変数を用いて評価する．情報の発信者は，受信者が有益と判断するメッセージの内容や，伝播メディアなど，メッセージ特徴の選択において，目標変数に影響を与えるコンテンツの要素を明らかにしたいという要望が高まっている．回帰に基づくメッセージ特徴では，利用者が有益と判断するコンテンツのメッセージ特徴を目標変数に基づいて評価する．

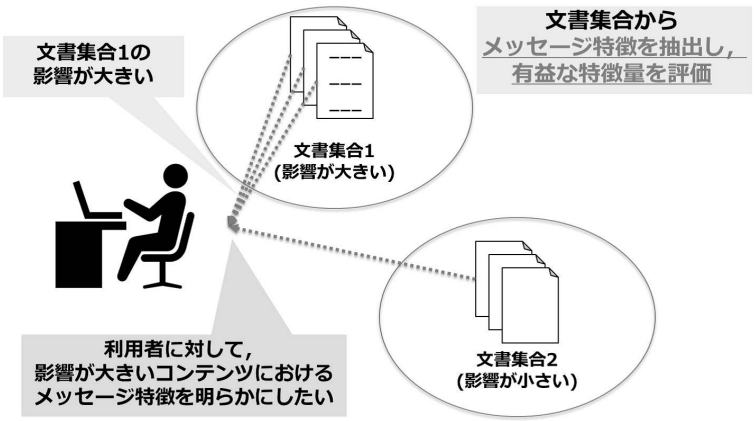

図 5.1: 目標変数への影響に関する問題

回帰に基づくメッセージ特徴は，特徴量の抽出，特徴量の変換，メッセージ特徴の評価で構成されている．回帰に基づくメッセージ特徴の評価実験では，既存研究の特徴量抽出手法を用いて，テキスト情報の特徴量を抽出する．特徴量の変換では，3.2.1 節にて示した非負値行列因子分解 (NMF) を用いて，メッセージ投稿となる変換特徴量を抽出する．そして，提案手法を用いて，特性に基づくメッセージ特徴を抽出する．メッセージ特徴は，複数の特徴量からなる特徴量の集合である．

評価実験で用いるテキスト情報は，4.2 節にて示した特徴量である．後述するが，5.3 節の評価実験では，4.2 節の特徴量に加え，辞書にプライバシー侵害に関連のある既存研究の辞書を用いた．したがって，5.3 節の実験の観測行列は合計値 2,073 である．5.2 節にて提案手法である NMF を適用した結果であるメッセージ特徴から，回帰に基づくメッセージ特徴を評価する方法を示す．実装方法は，4.2.4 節にて示した方法を用いる．そして，5.3 節および 5.4 節にて，提案手法の評価実験を示す．

5.2 回帰に基づくメッセージ特徴の特徴選択手法

NMF の係数行列 U の成分 $u_{m,i}$ は文書 i の基底 m への重みを表す．NMF で行列分解された基底行列 H は，観測行列 Y の特徴量の共起成分がグルーピングされた結果である．回帰に基づくメッセージ特徴の抽出方法では，入力特徴量である係数行列 U の成分 $u_{m,i}$ を変化させ，重要な基底を評価する．係数行列 U の各基底の列を除外し，回帰分析を行うことで，結果に影響がある特性に基づく基底 H が推定できる．

回帰に基づくメッセージ特徴の抽出では，非線形回帰手法[1]の一つであるサポートベクター回帰モデル (SVR : Support Vector Regression)[122] 用いる．SVR は，入力 $x_i \in R^p (i = 1, 2..., l)$ を特徴空間へ非線形写像し，特徴空間で線形回帰を行うモデルである．SVR は汎化能力が高い回帰モデルであることが知られている [124]．SVR の回帰関数を式 (5.1) に示す．

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x) + bias \tag{5.1}$$

$K(x_i, x)$ は入力 x_i を特徴空間へ写像するカーネル関数である．本研究では，RBF カーネル [124] を用いる．RBF カーネルを式 (5.2) に示す．

$$k(x_i, x) = \exp(\parallel x_i - x \parallel^2 / 2\sigma^2) \tag{5.2}$$

α_i,α_i^*,$bias$ などの詳細は文献 [124] などを参照していただきたい．

本研究の基底の評価では，各説明変数 x_i を順に除外した場合で，目標変数に対する各予測誤差を算出した．予測誤差が大きい入力の組み合わせは，影響大きい説明変数が除外された組み合わせである．したがって，説明変数 x_i から目標変数への影響を判別できる．

[1]高次の項も含む数式．対して線形 (一次) は，ベクトルの集合に対して，その要素の定数倍と加法で特徴づけられる数式．

5.3 - 実験 2-1 - 伝播メディアにおけるプライバシーの課題への応用

5.3.1 概要

スマートフォンなどの常時携帯型の端末が広く普及したことで，オンライン上でのコミュニケーション（SNS) が急速に広まっている．バイラル・マーケティング[2]などユーザ相互の口コミ情報などを SNS 上で拡散する手法やサービスの宣伝活動も認識されてきている．一方で，SNS など不特定多数による情報発信を基盤とするメディアでは，ふとしたことから，他者のプライバシーを侵害したり，誹謗中傷に加担してしまうことが問題となっている．特に，バイラル・マーケティングの過程などで，このような問題が発生すると，一個人の問題では済まなくなり，社会的な損失も無視できない規模に拡大することがある．

また，プライバシーに関しては， 2018 年 5 月 25 日に適用された EU の一般データ保護規則（General Data Protection Regulation: GDPR） がある．GDPR は，個人データの安全な管理措置と個人データの越境に関する規制であり，オンライン・サービスも含め，個人データの処理，および個人データを欧州経済領域 (European Economic Area : EEA) から第三国に移転するための法的要件の規定である．このように，近年，プライバシーに関わるデータの処理には，より慎重に扱い，かつ速やかな対応が求められている．“実験 2-1” では，オンライン・サービスの一つである SNS で投稿者が意図したか否かに関わらず，共有されたコンテンツが，他者に不愉快な感情を与えるコンテンツであるか．あるいは SNS で共有されたコンテンツが，他者がプライバシー侵害と判断するかを，コンテンツから判断できるかを論じる．

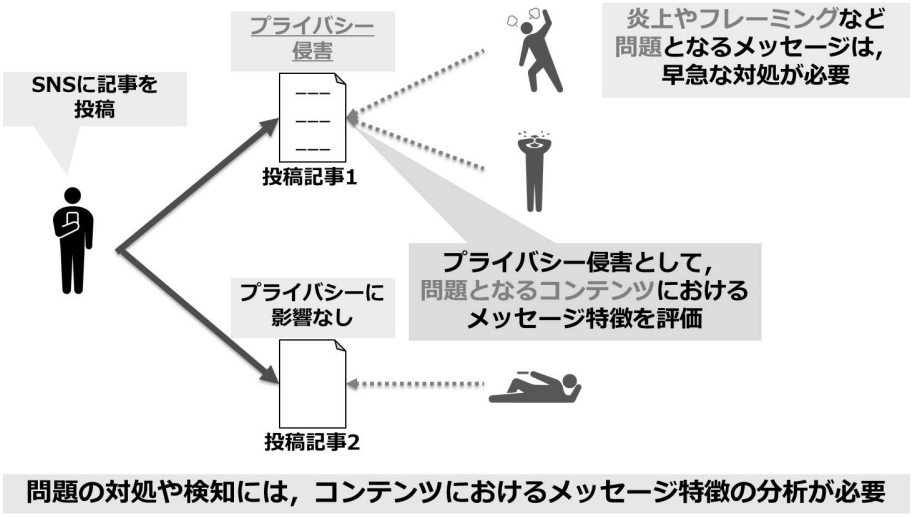

図 5.2: 研究背景 (実験 2-1)

[2]バイラルは「ウイルス (Virus)」の意，ウイルスが感染する様子に似ていることから，インターネットを媒体とした口コミを利用して，商品やサービスの宣伝を行う方法．オンライン・ショッピングで「この商品を友達に勧める」という選択肢を用意し，情報の伝播を促す手法．

“実験 2-1” では，プライバシー侵害のアンケート結果と既存研究の特徴量でプライバシー侵害に影響があるメッセージ特徴を明らかにする．“実験 2-1” のメッセージ特徴抽出手順の概要を図 5.3 に示す．図 5.3 の (1) から (4) が，“実験 2-1” のメッセージ特徴抽出手順を示すステップである．

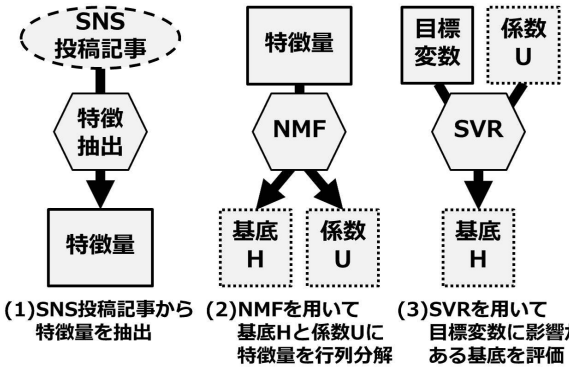

図 5.3: 実験 2-1 メッセージ特徴抽出手法

図 5.3(1) の特徴量の抽出では 4.2.1 節，4.2.2 節，4.2.3 節に示した特徴量を抽出している．加えて，“実験 2-1” では，プライバシー侵害に関連のある既存研究の辞書 [63][67] を用いる．抽出する特徴量は既存研究の特徴量を組み合わせた，33 個，各特徴量を水平方向に連結した際の次元数は 2,073 次元である．そして (2) から (3) が，“実験 2-1” のを評価する手順である．(2) では NMF を用いて行列分解を行い，特徴量の変換を行う．(2) および (3) のメッセージ特徴抽出の基底選択では，5.2 節の回帰による基底選択を用いた．

“実験 2-1” の目標変数には，プライバシー侵害のアンケート結果を用いた．“実験 2-1” は，2.3 節で示した従来研究と比較し，テキスト情報の特徴量を網羅的に抽出している．また，プライバシー侵害に関する従来研究で，NMF による行列分解で特徴量変換し，網羅的なテキスト情報から基底を評価する研究は著者の知る限り，行われていない．そこで，“実験 2-1” では，回帰に基づくメッセージ特徴を用いて，評価実験を行う．

5.3.2 評価対象

“実験 2-1” では，SNS の一つである Twitter より，プライバシー侵害に関連するハッシュタグを指定し，取得した SNS 投稿記事を評価する．ハッシュタグは「#写真，#集合写真，#隠し撮り，#盗撮，#無音カメラ，#バカ発見器」である．アンケート評価を行った 96 件が評価対象である．SNS 投稿記事より抽出する特徴量の次元数は 2,073 次元である．したがって，“実験 2-1” における NMF の観測行列 Y は，96 行 2,073 列の長方行列である．

“実験 2-1” のメッセージ特徴の抽出では，5.2 節の回帰による基底選択を用いている．“実験 2-1” の SVR の目的変数には，プライバシー侵害のアンケート結果を用いた．また，説明変数 x_i には，NMF の基底 H に対する重みである係数行列 U の値である．プライバシー侵害のアンケートでは，収集した Twitter の静止画付き投稿が，プライバシー侵害に該当しているかのアノテーションを実施した．アンケート実施人数は 55 名，収集対象としたハッシュタグは「#写真，#集合写真，#隠し撮り，#盗撮，#無音カメラ，#バカ発見器」である．したがって，“実験 2-1” では目標変数をアンケート評価の結果とし，アンケート評価の差異でプライバシー侵害に関連する SNS 投稿記事を判別し，そのメッセージ特徴を評価する．

5.3.3 評価方法

“実験 2-1” ではメッセージ特徴の評価に 3.3 節の分類器，Ada Boost，Random Forests，MLP，K-NN を用いる．分類器の評価では，プライバシー侵害に関連する SNS 投稿記事を 3.3.5 節の適合率，再現率，F-measure を算出し，評価する．分類器では，4.3 節の回帰に基づいたメッセージ特徴の結果である基底を用いる．分類評価では，基底の寄与率上位の特徴量を用いる．また，分類評価ではメッセージ特徴抽出に用いたプライバシー侵害アンケートと同様のハッシュタグのデータを抽出した．評価に用いたデータは 2008 年から 2012 年までの Twitter の SNS 投稿記事を基にしており，収集期間の投稿記事のデータ総数は約 12 億件である．評価に用いたデータを表 5.1 に示す．

表 5.1: プライバシーのメッセージ特徴分析に用いたデータ

特徴量名	次元数
評価対象の SNS 投稿記事	Twitter のツイートデータ
収集期間	2008 年 10 月 - 2012 年 5 月
	(但し，一部期間は欠落あり)
収集期間の投稿記事の総数	1,198,426,092 件 (約 12 億件)
抽出した画像付き投稿記事数	5,249,748 件 (約 520 万件)
評価対象タグの投稿記事数	423 件
評価対象タグ以外の投稿記事数	5,249,325 件

実験では収集期間の投稿記事より，ファイルの拡張子「.png」「.jpg」「.jpeg」「.gif」および「twimg」が含まれる投稿記事を抽出し，画像付き投稿記事としている．また，アンケート評価と同様の評価対象タグの投稿記事が「プライバシー侵害に該当した投稿記事」である．評価対象のタグは 5.3.2 節と同様に,「#写真，#集合写真，#隠し撮り，#盗撮，#無音カメラ，#バカ発見器」である．評価対象タグ以外の画像付き投稿記事は「プライバシー侵害に該当していない投稿記事」としてタグ付けを行った．実験では，評価対象タグ以外の画像付き投稿記事から，4 万件をランダムに抽出し，評価に用いた．“実験 2-1” では適合率，再現率，F-measure は層化 5 分割交差検証で算出した．また，“実験 2-1” では分類器による評価に加え，因果関係の推定を行う．因果関係の推定では，3.5 節のベイジアンネットワークを用いる．評価結果である基底の寄与率上位 100 個の特徴量を用いて，プライバシー侵害の判断に因果関係がある特徴量を明らかにする．プライバシー侵害は受信者の主観に基づくことから，一意的なラベル付は困難である．本研究におけるベイジアンネットワークでのプライバシー侵害の評価では，アンケート実施者の評価の平均値で，プライバシー侵害に「該当」,「該当しない」という 2 値の名義尺度に変換している．加えて，特徴量の変数の値は，中央値を基準に，基底への寄与率が「高い」,「低い」という 2 値の名義尺度に変換している．

5.3.4 基底選択の評価結果

提案手法の結果である基底 H を SVR で評価した結果を表 5.2 および図 5.4 に示す．結果，表 5.2 および図 5.4 より，最も MAE および RMSE の値が大きくなるのは，基底 3 の説明変数を除いて SVR を行った場合である．したがって，最もプライバシー侵害に影響が大きい基底は基底 3 である．同様に，基底 2 および基底 9 も影響が大きい基底であることが明らかになった．

表 5.2: 実験 2-1 における基底 H の評価

	MAE	RMSE
All	0.8627	1.1470
Remove 0	0.8602	1.1343
Remove 1	0.8388	1.0624
Remove 2	0.8775	1.1613
Remove 3	0.9091	1.1946
Remove 4	0.8506	1.1196
Remove 5	0.8380	1.1075
Remove 6	0.8395	1.0719
Remove 7	0.8493	1.1165
Remove 8	0.8204	1.0683
Remove 9	0.8672	1.1564

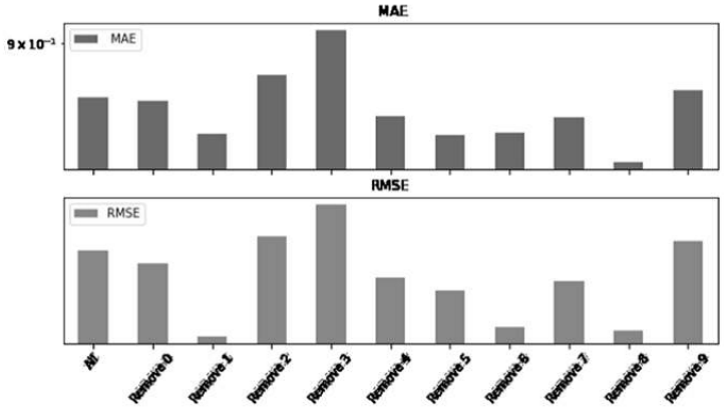

図 5.4: 実験 2-1 における基底 H の評価

各基底を特徴量の種別ごとにヒートマップを表したものを図 5.5 に示す．既存研究の特徴量は，メルヴィル・デューイの十進分類法 [133] に基づき，表層情報 (Surface Layer Information)，話題 (TOPIC)，語種 (Word Type)，基本語 (Basic Vocabulary)，意味属性 (Semantic Attributes)，言語表現 (Verbal Expression)，文末表現 (Sentence End Expression)，品詞 (Pos Type)，固有表現 (Unique Expression)，評価表現 (Evaluation Expression) に分類した．図 5.5 の色の濃さは，基底に対する種別の係数値が大きいことを示している．

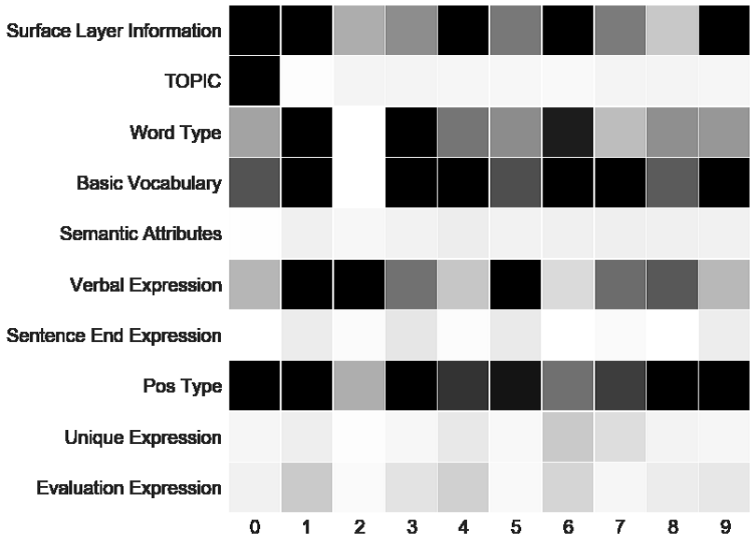

図 5.5: 実験 2-1 NMF 基底 H

図 5.5 から，基底 3 は語種，基本語，品詞の寄与率が高い．ここで，最も影響が大きい基底は基底 3 である．提案手法の結果，基底に基づく特徴量の上位 20 個を表 5.3 に示す．

寄与率による評価では，語種判定 (3)，文字種特徴量 (頻度 - カタカナ)，文字種特徴量 (比率 - カタカナ) など外来語の特徴量が寄与率が高い結果であった．また，評価値表現辞書や単語感情極性対応表 (ネガティブ単語頻度) など評価表現の特徴量も寄与率が高い結果となっている．加えて，従来法である文献 [63] の特徴量である LPSW などが寄与率上位に評価されている．したがって，基底 3 はプライバシーに影響が高い基底であると解釈できる．

表 5.3: 実験 2-1 における基底 H の評価

特徴量	
意味分類体語彙表 意味分類コード (1.12)	品詞特徴量 (副詞)
分類項目一覧表 意味分類コード (1.503)	評価値表現辞書
分類項目一覧表 意味分類コード (2.503)	単語感情極性対応表 (ネガティブ単語頻度)
分類項目一覧表 意味分類コード (3.503)	機能表現タグ (B-否定)
分類項目一覧表 基本語 (1)	文字種特徴量 (頻度 - カタカナ)
TOPIC(LSI)	文字種特徴量 (比率 - カタカナ)
日本語教育基本語彙 意味分類コード (1461)	機能表現タグ (B-順接限定)
意味分類体語彙表 意味分類コード (1.461)	LPSW
語種判定 (3)	機能表現タグ (I-否定)
日本語教育基本語彙 基本語 (0)	意味分類体語彙表 基本語 (*)
分類項目一覧表 意味分類コード (1.503)	品詞特徴量 (副詞)

5.3.5 分類実験による特徴量の評価結果

5.3.4 節で提案手法を適用し，得られたプライバシーの影響が大きい基底 3 の寄与率上位 100 個の特徴量を用いて分類器による基底の評価を実施した．分類器による分類結果を表 5.4 に示す.

表 5.4: 実験 2-1 分類器による評価

	適合率	再現率	F-measure
Ada Boost	0.88	0.85	0.87
Random Forests	0.90	0.64	0.75
MLP	0.87	0.72	0.78
K-NN	0.89	0.68	0.77

結果，すべての分類器で「適合率」の評価基準で高い結果が得られた．最も「適合率」が高い分類器は「Random Forests」の「0.90」である．一方で,「再現率」では「Ada Boost」が「0.85」，次いで「MLP」では「0.72」が得られた．結果，適合率と再現率の調和平均値である「F-measure」では,「Ada Boost」が最も高い「0.87」を示した．従来法である文献 [7] のプライバシーに関する投稿記事の抽出評価では,「F-measure」で最も良い評価結果が「0.73」である．したがって，従来法と比較しても良い評価結果が得られた．このため，提案手法によって得られた基底の特徴量を用いることで，プライバシー侵害に影響がある SNS 投稿記事の分類が行えることがデータで示された.

5.3.6 因果関係がある特徴量の評価結果

アンケート結果を目的変数として，ベイジアンネットワークを構築し，プライバシー侵害の変数ノードの原因ノードとなった特徴量を表 5.5 に示す．ベイジアンネットワークによる因果関係のある特徴量の推定では，5.3.4 節で評価した基底 3 に加えて，重み付け係数値上位の基底である基底 2，基底 9 の評価も行った．

表 5.5: プライバシー侵害記事の評価結果の特徴量

No.	基底 2	基底 3	基底 9
1	TOPIC(LSI)	機能表現タグ	TOPIC(LSI)
		(B-不許可)	
2	文字種	意味分類コード (2)	単語感情極性
	(比率 - カタカナ)	(1.564)	(ポジティブ比率)
3	語種判定 (3)	-	品詞特徴量 (動詞)
4	固有名詞 (一般)	-	-

提案手法で評価した基底 3 の特徴量の寄与率で因果関係を推定した際は，機能表現タグ (B-不許可) と，意味分類コード (2) の (1.564) がプライバシー侵害の原因として評価された．機能表現タグ (B-不許可) の辞書は「て，ちゃ」である．一方で，意味分類コード (2) の (1.564) は分類では「体の類自然物および自然現象」である．しかし，詳細な分類では「魚 (魚類円口類)」に該当し，辞書の単語は「魚，たい，あじ」などである．したがって，ベイジアンネットワークの推定では基底 3 とアンケート結果を用いて推定された辞書の特徴量がプライバシー侵害と因果関係があるとは十分に言えない．

一方で，係数値上位の基底 2 の特徴量の寄与率で因果関係を推定した場合，カタカナの比率や語種判定 (3) など，外来語の特徴量がプライバシー侵害の原因として評価された．推定された特徴量である語種は「ホテル・旅館・やどや」など，同じ意味であっても，イメージが異なる特徴量である [134]．推定結果である「語種判定 (3)」は外来語である．既存研究の評価では，外来語は「女性・ファッション」の要因が相対的に高い [135]．しかし，肯定的なイメージを付加する反面,「異質性，斬新な響き，ステイタス顕示」など否定的な印象で評価される要因もある [136]．また,「プライバシー」という単語も外来語である．したがって，基底 2 で評価された外来語はプライバシー侵害に関連し，否定的な印象で評価される外来語があると推察される．また，固有名詞も因果関係がある結果となっていた．

このため，特定の固有名詞がプライバシー侵害と関連していると推察される．基底 9 では動詞の比率がプライバシー侵害の原因として評価された．文書に動詞が多い場合,「行動，変化を述べる」などの動きの描写がある動的な記述の文章であるとされている [94]．また，動詞は既存研究のコーパス間比較では，論文，新聞と比較し，日記が比率が高い [76]．

したがって，行動，変化を述べる，日記のような個人の投稿がプライバシー侵害と関連していると推察される．基底 2，基底 9 では TOPIC(LSI) が原因として評価された．したがって，プライバシー侵害に該当する話題があると推察される．ここで，提案手法の結果，得られた基底は基底 3 である．しかし，ベイジアンネットワークの推定では，基底 3 では因果関係のある特徴量の推定は十分ではない．一方で，係数値上位の基底 2 および基底 9 では，因果関係があると推定される特徴量が得られた．結果，ベイジアンネットワークで NMF の基底から因果関係のある特徴量を推定する場合は，単一の基底でなく，複数基底の評価が必要であると推察される．

5.4 - 実験 2-2 -
伝播メディアのコミュニケーションに対する共感の課題への応用

5.4.1 概要

近年，情報通信技術の進歩により，CGM やオンラインコミュニティなどインターネットにおける個人を主体に情報発信を行うソーシャルメディアが台頭してきている．ソーシャルメディアでは，個人の胸中や購入した商品やサービスの所感などを自由に情報発信できる．したがって，情報収集や気楽なコミュニケーション，時間をかけた議論など目的に応じた多くの利点がある．

オンラインコミュニティでは，商品やサービスの利用者が，内容や使い勝手などの情報交換も盛んに行われている．オンラインコミュニティにおける情報推薦や情報の管理では，利用者がより良く利用できるために，コミュニティの質問記事を適切に分類することが求められている．

コミュニティなどソーシャルメディアにおける情報は，多くの読者やユーザーが閲覧するため，社会的な影響も大きい．ソーシャルメディアにおける発信に対する返信は，ポジティブあるいはネガティブに共感した場合に行われ，マクロ的[3] な秩序現象が創発される場合もある．返信は利用者にとって，共感，有益あるいは重要な場合にのみ行われる．そこで，提案手法では，質問記事の返信数に着目して，利用者が有益と判断するコンテンツの特徴量を明らかにする．有益と判断するコンテンツの特徴量が明らかになることで，オンラインコミュニティにおける共感の要素が明らかになる．共感の要素が明らかになることで，情報カスケードの分析や動向や周期性の評価が行える．

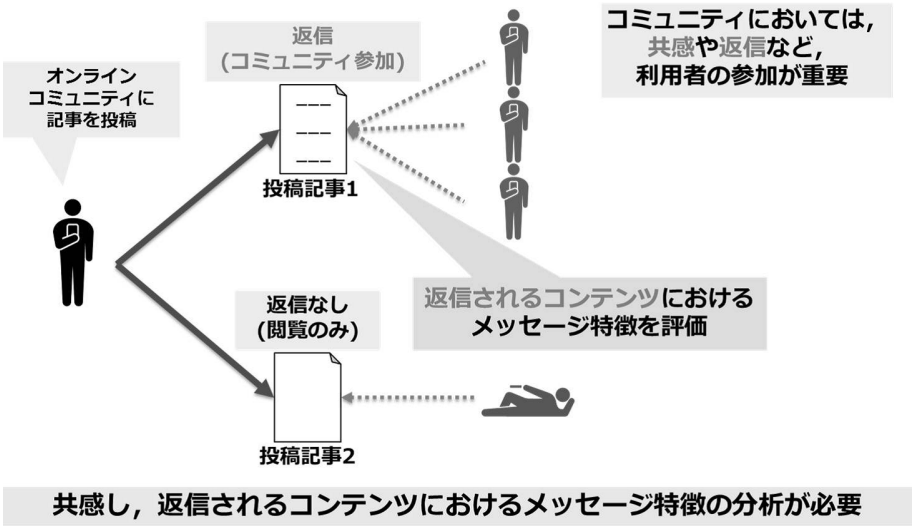

図 5.6: 研究背景 (実験 2-2)

[3]現象に対する視野が大きいさま．巨視的.

“実験 2-2” では，オンラインコミュニティの返信数と既存研究の特徴量でオンラインコミュニティの返信数に影響が大きい特徴量の集合であるメッセージ特徴を明らかにする．“実験 2-2” のメッセージ特徴抽出手順の概要を図 5.7 に示す．図 5.7 の (1) から (3) が，“実験 2-2” のメッセージ特徴抽出手順を示すステップである．

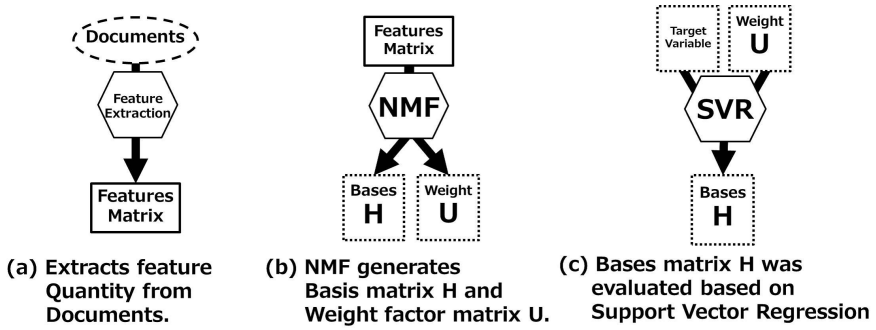

図 5.7: 実験 2-2 メッセージ特徴抽出手法

図 5.7(1) の特徴量の抽出では 4.2.1 節，4.2.2 節，4.2.3 節に示した特徴量を抽出している．抽出する特徴量は既存研究の特徴量を組み合わせた，31 個，各特徴量を水平方向に連結した際の次元数は 2,071 次元である．そして (2) から (3) が，“実験 2-2” のを評価する手順である．(2) では NMF を用いて行列分解を行い，特徴量の変換を行う．(2) および (3) のメッセージ特徴抽出の基底選択では，5.2 節の回帰による基底選択を用いた．節で後述するが，“実験 2-2” の目標変数には，オンラインコミュニティの返信数を用いた．“実験 2-2” は，オンラインコミュニティの質問記事の返信数を評価している 2.1 節で示した従来研究と比較し，テキスト情報の特徴量を網羅的に抽出している．また，従来研究で，NMF による行列分解で特徴量変換し，網羅的なテキスト情報から基底を評価する研究は著者の知る限り，行われていない．加えて，“実験 2-2” では，行列分解の結果得られる基底を係数値の重み付けだけでなく，SVR を用いて，質問記事の返信数に影響が大きい基底を評価した．次に “実験 2-2” の実験環境について述べる．

5.4.2 評価対象

伝播メディアに用いるデータセットの簡単な説明を行う．本研究では，コミュニティの返信を特徴付ける有益なコンテンツに特有の基底を目的に，2 種類のオンラインコミュニティを用いて評価する．コミュニティはゼネラル・メディアやクラス・メディア [2] など，メディアの特性が大きく影響する．オンラインコミュニティ1 のデータセット 1 は Apple Inc.[4]が提供している Apple サポートコミュニティ[137] に 2008 年 10 月 1 日から 2014 年 1 月 24 日に投稿された質問記事 10,391 件を評価対象に用いる．オンラインコミュニティ2 のデータセット 2 は Stack Exchange, Inc. [5]が提供している 2018 年 9 月 2 日までに投稿された Stack Exchange Data Dump [138] のうち，stackover flow のデータ・セットである ja.stackoverflow.com の質問記事 35,945 件を評価対象に用いる．コミュニティの利用者の性質から，データセット 1 は社会全般の人々を対象としたゼネラル・メディア，データセット 2 は特定の集団などを対象とするクラス・メディアであると推定される．閲覧数に対する返信数の相関係数を算出した結果，データセット 1 では 0.42，データセット 2 では 0.76 という相関係数が得られた．相関係数と推定したメディアの性質から，本研究においては，コミュニティ2 における質問記事を返信される有益なコンテンツであると定義した．2 種類のデータセットを用いて，コミュニティの返信を特徴付ける有益なコンテンツに特有の特徴を抽出する．特徴抽出の評価実験では，2 種類のデータセットをマージさせ，観測行列とし，提案手法の適用および評価を行う．ここで，特徴量の次元数 K は，表層情報の次元数 22，アルゴリズムの次元数 600，辞書の次元数 1,449 の合計値 2,071 である．また，文書数 N は 46,336 である．したがって観測行列 Y は 46,336 行，2,071 列である．

本研究では，ソーシャルメディアで発信した情報を適切に分類し，情報推薦や情報収集など多種多様に利用することを目的としている．ゆえに，評価実験では，得られた特徴を他の伝播メディアを用いた文書分類で評価する．伝播メディアの文書分類では，4 種類のデータセットを用いる．データセット 3 は青空文庫[6]の電子書籍のテキストデータ 14,266 件を評価対象に用いる [139]．データセット 4 は，Twitter [7]のツイートデータを用いる．収集期間は 2008 年から 2012 年までで，投稿記事のデータ総数は約 12 億件から，画像付き投稿記事を 40,000 件をランダムに抽出し，評価対象に用いる．画像のファイルの拡張子は,「.png」「.jpg」「.jpeg」「.gif」および「twimg」が含まれる投稿記事を対象とした．データセット 5 は日本語 Wikipedia[8]が提供している Wikipedia のデータベース・ダンプから 40,000 件をランダムに抽出し，評価に用いる [140]．データセット 6 は株式会社ロンウイット[9]が提供している NHN Japan 株式会社のニュースサイト「livedoor ニュース」の記事データ 7,367 件を評価対象に用いる [141]．伝播メディアの評価実験では，4 種類のデータセットをマージさせ，文書分類実験を行う．したがって，文書分類の実験を行う文書数は 101,633 である．

[4]Apple : https://www.apple.com
[5]Stack Exchange: Hot Questions : https://stackexchange.com/
[6]Aozora Bunko : https://www.aozora.gr.jp
[7]Twitter : https://twitter.com/
[8]Wikipedia : https://ja.wikipedia.org
[9]RONDHUIT : https://www.rondhuit.com/

5.5 基底選択の評価結果

NMF の基底数を M = 50 に設定し，提案手法を適用し，基底の評価を行った．結果を図 5.8 に示す．

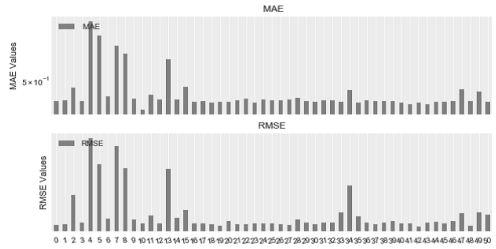

図 5.8: 返信数が目標変数の場合に影響が大きい基底

横軸が基底 m であり，縦軸が目標変数である返信数の予測誤差である．結果，基底 4 が非負値で最も返信数に影響を与える基底であると評価された．

5.5.1 非線形回帰を用いた特徴量選択の評価

提案手法の結果得られた特徴から，特徴量選択を用いた評価を行う．特徴量選択では，提案手法で得られた特徴である基底に対して，寄与率の高い特徴量を特徴量選択し，評価した．評価では SVR を用いて，回帰分析で評価した．回帰分析の目標変数は，返信数である．評価基準は，目標変数との予測誤差である．予測誤差には，MAE を用いた．結果を図 5.10 に示す．図 5.10 から，限定的だが，回帰分析においては特徴量選択は，返信数の予測誤差減少に有効であることが明らかになった．最も精度が良い特徴量の選択数は，50 個であった．

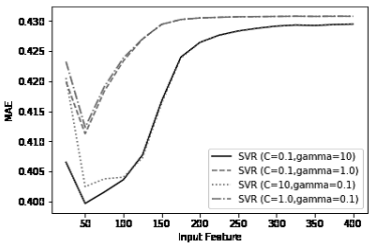

図 5.9: 実験 2-2 における基底選択の結果を用いた MAE(1)

5.5.2 分類器を用いた特徴量選択の評価

次に，回帰分析と同様に，文書分類を用いて特徴量選択を用いた評価を行う．特徴量選択では，提案手法で得られた特徴である基底に対して，寄与率の高い特徴量を特徴量選択し，

評価した．分類器には，AdaBoost，RandomForests，Multi-layer Perceptron(MLP)，K-Nearest Neighbors(K-NN) を用いる．結果を表 5.6 に示す．一行目の数値は選択基底における特徴量の選択数である．表中の数値は文書分類の分類精度である．

表 5.6: 実験 2-2 における特徴量選択数に基づく文書分類結果 (F-measure Total)

Classifier	50	100	150	200	250	300
AdaBoost	0.88	0.92	0.96	0.97	0.97	0.97
RandomForests	0.85	0.89	0.91	0.93	0.93	0.94
MLP	0.90	0.94	0.95	0.95	0.95	0.94
K-NN	0.81	0.86	0.85	0.84	0.85	0.84

表 5.6 から，選択基底における分類精度は分類器で異なるが，F-measure 基準においては，100 個から 150 個で，特徴量選択数は，100 個から 150 個で，分類精度の増加が収束した．したがって，特徴量選択数は 100 個から 150 個が妥当であると言える．特徴量選択数が 100 個の場合の Precision Recall カーブ，受信者動作特性 (ROC)，AUC を示す．AUC は，ROC のカーブ下の領域である．

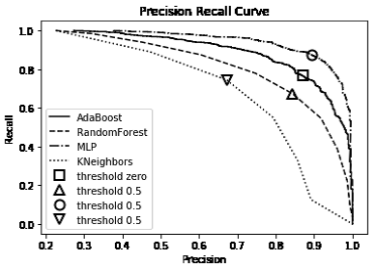

図 5.10: 実験 2-2 Precision Recall カーブ

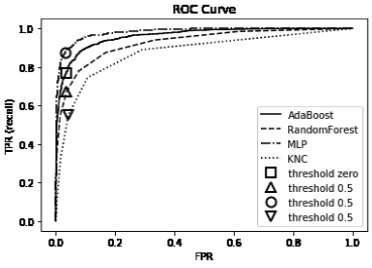

図 5.11: 実験 2-2 受信者動作特性 (ROC)

表 5.7: 実験 2-2 における AUC

Classifier	AUC
AdaBoost	0.9617
RandomForests	0.9271
MLP	0.9823
K-NN	0.8769

図 5.10 の分類スレッショルドの最適化を行うことで，適合率や再現率，F 値などの最適化が行える特徴量選択であることが明らかになった．同様に，図 5.11 の結果から，各分類器における FPR のスレッショルドの最適化で，TPR は向上することも明らかになった．したがって，提案手法で得られた特徴は，コミュニティの返信を特徴付ける有益なコンテンツに特有の特徴であると言える．

5.6 分類器を用いた複数基底の評価

提案手法では，目標変数に基づいて，返信数に最も影響が大きい基底を評価した．このセクションでは，影響が大きい基底を複数評価し，コミュニティ分類に影響が大きい他の基底を考察する．提案手法の結果得られた上位 4 つの基底を基にした，コミュニティ分類実験の結果を表 5.8 に示す．各基底における特徴量の選択数は 100 個である．表 5.8 中の数値は文書分類の分類精度である．

表 5.8: MAE 基準 複数基底評価 (F-measure Total)

Classifier	Base 4	Base 5	Base 7	Base 8
AdaBoost	0.92	0.98	0.94	0.82
RandomForests	0.89	0.97	0.91	0.80
MLP	0.94	0.98	0.96	0.83
K-NN	0.86	0.90	0.87	0.79

表 5.8 から，基底 4，基底 5，基底 7 は分類精度が高いことが明らかである．したがって，提案手法で得られた基底は，分類精度が高い基底の一つであると言える．一方で，基底 8 においては，上位 3 個の基底と比較した場合は，分類精度は十分ではない．したがって，結果から，妥当な基底を特徴した場合と妥当でない基底を特徴とした場合では，分類精度に明らかに差がでたと言える．このため，提案手法で得られた特徴は，妥当な特徴であり，提案手法は有効な特徴の選択手法であると言える．

一方で，結果から，提案手法で得られた基底 4 よりも，基底 5 や基底 7 の分類精度が高い結果となった．これは，文書分類はコミュニティの分類結果であり，基底選択においては，返

信数を基準として用いていることが起因していると言える．したがって，MAE の誤差を算出目標変数を妥当な変数とすることで，予測誤差が最大とする基底を選択することが，文書分類においても最大の精度を得られる結果とすることが期待できる．

5.7 選択特徴量を用いた伝播メディアの分類評価

このセクションでは，提案手法で得られた特徴を用いて，伝播メディアの文書分類評価を行う．伝播メディアの分類では，節 5.4.2 で示した Aozora Library，Twitter Tweet Data，Wikipedia Article，Livedoor News を用いる．複数の分類器を用いた，評価実験の結果を，表 5.9 から表 5.12 に示す．表中の数値は文書分類の分類精度である．分類実験に使用した特徴量の選択数は 100 個である．

表 5.9: Classification(AdaBoost)

Verbreitungsmedien	Precision	Recall	F-measure
Aozora Library	0.96	0.94	0.95
Twitter Tweet Data	0.91	0.75	0.82
Wikipedia Article	0.80	0.95	0.87
Livedoor News	1.00	0.99	0.99

表 5.10: Classification(Random Forests)

Verbreitungsmedien	Precision	Recall	F-measure
Aozora Library	0.98	0.99	0.99
Twitter Tweet Data	0.95	0.90	0.92
Wikipedia Article	0.91	0.96	0.93
Livedoor News	1.00	1.00	1.00

表 5.11: Classification(Multi-layer Perceptron)

Verbreitungsmedien	Precision	Recall	F-measure
Aozora Library	0.96	0.99	0.97
Twitter Tweet Data	0.92	0.76	0.83
Wikipedia Article	0.81	0.97	0.88
Livedoor News	0.93	0.78	0.85

結果，各分類器で分類精度に差はあるが，得られた特徴は，伝播メディアの文書分類に有効であることが明らかになった．最も精度の良い分類器である RandomForests を用いること

表 5.12: Classification(K-Nearest Neighbors)

Verbreitungsmedien	Precision	Recall	F-measure
Aozora Library	0.93	0.87	0.90
Twitter Tweet Data	0.80	0.71	0.75
Wikipedia Article	0.75	0.91	0.82
Livedoor News	0.83	0.50	0.62

で，F-measure 基準で，すべての伝播メディアで，分類精度 0.9 以上の優れた分類精度を示した．したがって，提案手法で得られた特徴は，他のデータ・セットの文書分類にも用いることができる有効な特徴であると言える．

5.8 むすび

本章では，非負値行列因子分解アルゴリズムから得られるメッセージ特徴を，サポートベクター回帰モデルと組み合わせることで，伝播メディアにおける目標変数への影響に関する問題の改善が期待できるメッセージ特徴の選択手法を提案した．

そして，選択手法で得られたメッセージ特徴を伝播メディアにおけるプライバシーの課題および伝播メディアのコミュニケーションに対する共感の課題で，性能評価実験を行った．伝播メディアにおけるプライバシーの評価では，プライバシー侵害のアンケート結果と SNS 投稿記事から，メッセージ特徴を評価した．また，伝播メディアのコミュニケーションに対する共感の評価では，オンラインコミュニティの返信数に基づいたメッセージ特徴を評価した．

以下，本章で得られた結果を示す．

- 伝播メディアにおけるプライバシーの課題に対する性能評価実験では，2,000 次元を越える高次元の特徴量から有効なメッセージ特徴を評価した．結果，メッセージ特徴に基づいた特徴量で，プライバシー侵害の投稿記事と，プライバシー侵害でない投稿記事の優れた分類精度結果が得られた．また，因果関係推定においても，複数基底の評価で，有効な結果が得られた．したがって，提案手法は，伝播メディアにおけるプライバシーの課題において，有効なメッセージ特徴の選択手法であると言える．

- 伝播メディアのコミュニケーションに対する共感の課題に対する性能評価実験においても，2,000 次元を越える高次元の特徴量から有効なメッセージ特徴を評価した．結果，メッセージ特徴に基づいた特徴量で，コミュニティ分類において，優れた分類精度が得られた．また，複数の基底評価を行った場合においても，提案手法で得られた結果は優れた結果が得られた．加えて，返信数の予測においても返信数の予測誤差の減少に有効な結果が得られた．したがって，提案手法は，伝播メディアのコミュニケーションに対する共感の課題において，有効なメッセージ特徴の選択手法であると言える．

- 伝播メディアにおける目標変数への影響に関する問題において，2 種類の評価実験で，優れた分類精度を示したことから，本研究におけるメッセージ特徴の有効性，また，選択手法においても優れた結果が得られたと言える．

- 伝播メディアのコミュニケーションに対する共感の課題に対する性能評価実験において，得られた特徴量で，異なる伝播メディアの文書分類を行った場合においても，優れた分類精度が得られた．したがって，伝播メディア分類のための特徴量選択手法とした場合においても，優れた結果が得られたと言える．

ここで，回帰に基づくメッセージ特徴の特徴選択手法では，2 種類の伝播メディアにおける課題にて，性能評価実験を行った．本研究においては，テキストコミュニケーションメッセージで性能評価実験を行った．評価実験の結果から異なるテキストコミュニケーションメッセージにおいても適用可能であり，伝播メディアをはじめとするテキストコミュニケーションにおいて，幅広い応用が期待できる．

第6章　結論

6.1　総括

本論文は，伝播メディアにおけるコミュニケーションにおいて情報伝達されるメッセージに関するメッセージ特徴の考察，メッセージ特徴の特徴選択手法，および伝播メディアの課題への適用可能性という主眼[1]に研究した結果をまとめたものである．

伝播メディアでは，文書データ，画像データ，音声データだけでなく多くの特徴量が存在する．本論文では，伝播メディアの特徴量の種類や次元数に対応可能な非負値行列因子分解アルゴリズムに着目し，有効な構成成分であるメッセージ特徴を特徴量選択に応用する手法を提案した．非負値行列因子分解アルゴリズムは，信号処理などでも用いられており，適用分野に応じて損失関数やアルゴリズムの選択肢，またアルゴリズムにおいても，改善が進められている優れたアルゴリズムである．優れたメッセージ特徴が明らかになることで，不要なメッセージ特徴を抑制し，優れたメッセージ特徴でコミュニケーションすることが期待できる．

本研究における提案手法として，構成成分であるメッセージ特徴を課題に応じて選択する手法を提案した．提案手法のアプローチでは，メッセージ特徴の 2 種類の選択手法を提案した．

まず第一のメッセージ特徴の選択手法では，非負値行列因子分解アルゴリズムとグレゴリー・ベイトソンの情報の定義である「“違い” を生む “違い”」を組み合わせた．本研究では，第一のメッセージ特徴の選択手法を差異に基づくメッセージ特徴の特徴選択手法とした．メッセージの話題性，メッセージの平易化の性能評価実験を実施した．次に，第二のメッセージ特徴の選択手法では，非負値行列因子分解アルゴリズムとサポートベクター回帰モデルを組み合わせた．本研究では，第二のメッセージ特徴の選択手法を回帰に基づくメッセージ特徴の特徴選択手法とした．メッセージにおけるプライバシー，メッセージに対する共感の性能評価実験を実施した．性能評価実験で用いた伝播メディアの課題は，マス・コミュニケーションにおける 3 要素である事実の報道，解説・啓発，娯楽にプライバシーの要素を含めた 4 要素に基づく課題で性能評価実験を行った．第一のメッセージ特徴の選択手法および第二のメッセージ特徴の選択手法のいずれにおいても，メッセージ特徴に基づいた特徴量で，優れた分類精度が得られた．したがって，非負値行列因子分解アルゴリズムから得られるメッセージ特徴を特徴量選択に応用する手法は，有効な提案手法であることが結果で示された．また，本研究の提案手法は，NMF における係数値の重みおよび SVR を用いている．したがって，評価実験のデータに依存していない．ゆえに，たとえ評価実験の途中で，評価対象のデータセットが変化あるいは欠損が発生した場合においても，提案手法を適用可能である．

[1]物事の最も重要な点．かなめ．眼目．

本研究では，提案手法の有効性を近年において主軸メディアとなったテキストコミュニケーションに着目し，性能評価実験を行った．設定要件として，設定した到達点は，受信者に有益と判断されるメッセージを判別できること，目標変数に影響が大きいメッセージを判別できること，メッセージ特徴が，伝播メディアの課題解決に有効であることである．メッセージ特徴を用いた伝播メディア課題への応用で，到達点を達成し，有効性が明らかとなった．4つの伝播メディア課題への応用結果を示す．

第一に，伝播メディアにおける話題性の課題においては，差異に基づくメッセージ特徴の特徴選択手法で，クラス・メディアのオンラインコミュニティにおいて優れた分類精度結果が得られた．また，非線形回帰においても，クラス・メディアのオンラインコミュニティにおいて，明瞭な閲覧数の予測精度の向上が得られた．特徴量選択手法として比較した場合においても，分類器依存せずに優れた分類精度が得られた．第二に，伝播メディアにおけるコミュニケーションの平易化の課題においては，差異に基づくメッセージ特徴の特徴選択手法で，平易化テキストの優れた分類精度が得られた．また，因果関係分析では，有益な結果が得られた．第一および第二の性能評価実験の結果から，差異に基づくメッセージ特徴においては，2種類の性能評価実験を実施し，有効な結果を確認できた．したがって，受信者に有益と判断されるメッセージの判別において，有効な結果が得られ，設定要件を達成できたと言える．

第三に，伝播メディアにおけるプライバシーの課題において，回帰に基づくメッセージ特徴の特徴選択手法で，プライバシー侵害でない投稿記事の優れた分類精度結果が得られた．また，因果関係推定においても，複数基底の評価で，有益な結果が得られた．第四に，伝播メディアのコミュニケーションに対する共感の課題において，回帰に基づくメッセージ特徴の特徴選択手法で，コミュニティ分類において優れた分類精度が得られた．また，複数の基底評価においても，有益な結果が得られた．加えて，返信数の予測においても返信数の予測誤差の減少に有効な結果が得られた．さらに，伝播メディア分類に応用した場合においても，有効な結果が得られた．第三および第四の性能評価実験の結果から，回帰に基づくメッセージ特徴においては，2種類の性能評価実験を実施し，有効な結果を確認できた．したがって，目標変数に影響が大きいメッセージの判別において，有効な結果が得られ，設定要件を達成できたと言える．

そして，4つの伝播メディア課題への応用結果から，非負値行列因子分解アルゴリズムから得られたメッセージ特徴で，いずれの課題においても，有効な結果を示すことができたと言える．したがって，メッセージ特徴が，4種類の異なる伝播メディアの課題解決に有効であり，テキストコミュニケーションに基づく数多くの伝播メディアにおいて，有効なメッセージ特徴を評価することができる優れた手法であると言える．このため，メッセージ特徴を用いた伝播メディアの課題解決に有効な結果が得られ，設定要件を達成できたと言える．

性能評価実験から，提案手法は2,000次元を越える高次元データにおいても有効であった．したがって，特徴量が変化あるいは増減した場合においても有効な優れた手法であると言える．ゆえに，今後，テキストコミュニケーションにおける異なる伝播メディアの課題解決においても有益な結果が得られることが期待できる．

6.2 今後の課題

本論文では，非負値行列因子分解アルゴリズムに基づくメッセージ特徴と，メッセージ特徴の選択手法，伝播メディア課題への応用について議論した．本節では，今後の課題として，3 点を示したい．

1 点目は，性能評価実験である．まず，非負値行列因子分解アルゴリズムは，非常に適応性が高いアルゴリズムである．本研究における提案手法は，特徴量が異なる場合においても，類似問題は，解析対象を観測ベクトルを並べた行列に表現することで適応できる．今後は，他言語を対象としたテキストコミュニケーションの評性能評価実験，画像データ，音声データ，信号処理など，伝播メディアにおける他の特徴量を用いた性能評価実験を検討する予定である．また，既存研究において特徴量変換における非負値行列因子分解アルゴリズムの優位性は示されているが，伝播メディアのメッセージにおける特徴量変換の優位性を明らかにする性能評価実験などを検討する予定である．加えて，伝播メディアのメッセージが変化した場合における受信者側の評価や推移の変化などについて検討する予定である．性能評価実験を行うことで，幅広い適応性を示し，情報社会のさらなる発展と問題解決に寄与したい．

2 点目は，メッセージの特徴量である．本研究では，伝播メディアのコミュニケーションに基づいて，メッセージに着目した課題解決を行った．結果から提案手法は，コミュニケーションのメッセージにおいては，解析対象を観測ベクトルを並べた行列に表現することで広く適応できることが期待できる一方で，コミュニケーションから特徴量が抽出できなければ，性能評価実験が行えない．したがって，提案手法が性能評価実験で，貢献を敷衍 (ふえん) [2]できる範囲は，送信者から受信者に対するコミュニケーションであり，かつメッセージから特徴量が抽出できる範囲である．一方で，言語コミュニケーション，記号非言語コミュニケーション，非言語コミュニケーションなどで用いられるメッセージには，非常に多くのコミュニケーション方法が存在する．感性システムに対して伝達できるメッセージには制約があり，コミュニケーションにおける構造や通信路にも，時間や計算量，伝搬量などの制約がある．サイバネティクス[3]などの感覚情報処理では眼，耳，鼻，舌，皮膚など五感の感覚器官により収集する [142]．しかし，サブリミナル[4]のような入出力処理は存在する．このため，非言語コミュニケーションやサブリミナル効果を評価対象とする場合は，コミュニケーションにおけるメッセージから，特徴量を抽出することが重要な課題である．特徴量の抽出方法を明らかにすることで，性能評価実験を行い，伝播メディアの課題解決に寄与したい．

3 点目はアルゴリズムの改善である．本研究においては，メッセージ特徴の抽出に非負値行列因子分解アルゴリズムを用いた．非負値行列因子分解アルゴリズムは，拡張モデルの提案や，アルゴリズムの改善，基底数などのパラメータ推論やモデルの複雑度の推論も進められている [112]．

性能評価実験，特徴量抽出，アルゴリズムの改善などについては今後の研究に期待したい．

---

[2]意義・意味をおし広げて詳しく述べること．たとえなどを用いてやさしく述べること．衍 (えん) は広げる意．

[3]生物と機械の間に共通点を見いだし，通信と制御の問題を統一的・体系的に追求する学問

[4]識閾 (しきいき) 下の意．意識されない，刺激に対して起きる知覚反応．通常の視覚・聴覚では捉えられない速度・音量によるメッセージを隠し，それを繰り返し流すことにより，潜在意識に働きかけること．サブリミナル効果は存在しているにもかかわらず知覚できない刺激が人に与える効果のこと [2]．

6.3 展望

本節では，本論文の成果を踏まえた望ましい発展の仕方と発展の方向について，私見を踏まえた展望を示したい．ここでは，適応性，発信者と受信者，後続のコミュニケーション，結果に基づく発展について示す．

一点目が，適応性としての発展である．2019 年までの発展において，情報アクセス技術は，急速な進歩を遂げた．常時携帯型の端末は広く普及し，1985 年にショルダーフォンだった携帯電話は，2013 年にはスマートフォンとなり年間 19 億台以上が販売されている [143]．一方で固定電話の全契約数や公衆電話施設構成数は緩やかに減少している [144]．新しい情報アクセス技術のトレンドは日に日に進歩し，ハイプ・サイクル [18] [5] など戦略的なテクノロジ・トレンド (Strategic Technology Trends) は変化し続けている．通信の分野では，最新の 4G システムはピーク速度が 10Gbit/s に達しようとしている．モバイル通信システムは約 10 年周期で世代交代が進むとされている．次世代の 5G システムでは 10〜20Gbit/s へ達する研究開発が行われている [145]．また，光ファイバ伝送の物理的な伝送限界は約 100Tbit/s とされているが，ペタビット級への大容量化技術も盛んに研究が進められている．通信システムは現状，低コスト化技術などが課題とされているが，今後も通信速度は高速化する傾向がある．移動体通信で扱うデータ量は年率 50%以上の勢いで増加することが予想されており，常時携帯型の基盤となる高周波集積回路技術の発展とワイヤレスインターネット社会の傾向は高まる [146]．さらに，世界のビックデータ市場は，ハードウェアの発展に加え，大量のデータをコンピュータで扱うことが容易になった要因もあり，年平均 11.7%の二桁成長が続いている [147]．

ここで，環世界 (Umwelt) という概念がある．環世界は客観的に存在する世界ではなく，生物ごとの認知能力によって生きられる世界である [148]．動物に与えられる環境刺激の意味のあり方は動物の種で異なる [149]．動物の感覚はそれぞれの種が固有に持つ個々の感覚器官に特有の仕方で与えられているためである．フォン・ベルタランフィ[6]が提示する個体性[7]アーサー・ケストラー[8]が提示する個体認識が相対的なものであるとする概念もある [150]．したがって，メッセージを評価する場合，動物や個体が感覚器官で知覚できる必要がある．

サブリミナルなど存在しているが知覚できない刺激は存在する．メッセージから特徴量が抽出できない場合，知覚できずに情報伝達や制御が行われる．一方で，特徴量が抽出できれば，ブラインド信号処理技術などで，メッセージを観測信号として，現信号を復元，分離再生を行うことができる [151]．一連の研究結果から，類似問題は解析対象を観測ベクトルを並べた行列に表現できれば原理的には解けると言える．非負値行列因子分解アルゴリズムは，画像データ，音声データ，信号処理など高い適応性を持つ．提案手法を用いた成果が，多くの性能評価実験に基づいた，適応的な発展であることを期待したい．

---

[5] 新しいテクノロジの発展と時間経過を示す先進テクノロジの成熟度と採用率のグラフ.
https://www.gartner.com/en/research/methodologies/gartner-hype-cycle

[6] カナダ・アメリカの理論生物学者 (1901-1972)．諸科学の方法論的統合を目指す「一般システム理論」を展開.

[7] 個別的存在だと経験したときだけは，個体性を直接に感じとれるが，周囲の生物に関しては，厳密に定義するわけにはゆかないとしている [150].

[8] イギリスの作家・ジャーナリスト (1905-1983)．全体からみると一部分だが，それ自体は完全なものとしての機能を持つものであるギリシャ語の $holos$ に由来した，ホロンの概念を提唱.

二点目が，発信者と受信者としての発展である．現状において，メディアのコミュニケーションは受信者に訴求するよう，刺激的な表現や欲求に訴えるような表現が多く用いられている．また今後，より刺激的な表現や，刺激の度合いの変化なども起きるであろう．メッセージ特徴は，マス・コミュニケーションの性質から，受信者に訴求するメッセージを作成する．

ヒトは感性システムである．影響力を行使するヒトに対しては，ミクロな引き込みやマクロな秩序，また環境から様々な刺激による擾乱 (じょうらん) が発生する．複数の発信者からメッセージとして，事前に調査を行った上で，想起[9]を目的とする類似記号，受信者に間接的に関連する記号などを用いる場合がある．類似記号は類似記号であり，対象の類似である．雑誌における画像などは，対象の類似でしかなく，実際にはインクによって作成された類似の形状や色彩を持つ類似記号ある．伝播メディアの発信者はメッセージとして，類似記号や電気機械信号などを送信し，行動や意志を制限，統制・束縛することも可能である．刺激による擾乱は，眼に対する刺激だけでなく，音声，電気機械信号などの刺激の場合もある．刺激を含めたメッセージは受信者にとっては，重要なメッセージである場合もあれば，雑音に相当する場合もある．提案手法で優れたメッセージ特徴は得られるが，メッセージを作成するのは，情報源である発信者である．集団的な社会意識は幅を効かせやすい．公的な制度や法システムの場合も同様である．今後も発信力のあるオピニオンリーダー[10]や影響力を行使する発信者は，制約の範囲内において，メッセージを受信，重要なメッセージの判別を行い，メッセージを作成する．公的コミュニケーションが私的コミュニケーションの領域に介入した場合，私的コミュニケーションが行われなくなり，後続するコミュニケーションが生成されない．また，社会的な制限や自律システムがより上位の自律システムから受ける制限が強くなった場合も同様である．発信者の立場に立つ場合は，課題解決に有効であるという観点に加えて，受信者に適切であるかなどを検討する必要も出てくるであろう．

受信者側の立場にたつ場合は，発信者側の立場を理解し，自身が，行動や意志を統制，あるいは伝播メディアとして用いられることを前提とした上で，メディアのコミュニケーションを受信し，行動する必要も出てくるであろう．紀元前にアリストテレス[11]のニコマコス倫理学[12]において,「いかなる技術，いかなる研究も，また，いかなる実践や選択も，ことごとく何らかのアガトン[13]を希求していると考えられる」という意思決定に関する表現が用いられている [1]．提案手法を用いた成果が，課題に対する効果に加えて，発信者と受信者の複眼的な観点で考察する，集合や共同体の分析に発展していくと期待したい．

---

[9]事象やそれに関する記憶心像を再現する過程．ギリシャ語の $anamnēsis$ であり，プラトンの用語．人間の魂が真の知識であるイデアを得る過程．イデアはプラトンの哲学で，感覚を超えた理性だけが認識できる時空を超えた永遠不滅の実在．

[10]ある集団の意思決定を方向づける人．構成員の意思や行動に影響を持つ人物．マスコミ理論の場合は第 1 次集団内の意見指導者．

[11]古代ギリシャの哲学者 (前 384-前 322)，プラトンの弟子，目に見える姿である形相 (エイドス) は現実の個物において内在・実現されるとし，あらゆる存在を説明する古代で最大の学的体系を立てた．

[12]アリストテレスの実践学．人間のなす事柄に関する哲学のうち，エートス (性格・習性など，個人の持続的な特質) を対象とした倫理学に関する著作の一つ．

[13]「善」またはそのもたらすところの「幸福」や「健康」をも意味したギリシャ語，$agathon$，ソクラテス以来，倫理学の中心課題とされた．

三点目が，後続のコミュニケーションとしての発展である．インターネットは誰もが閲覧できる環境であり，不特定多数のユーザが利用している．インターネットにおける共同体は，好感を持つ人と，友好的な関係を成立するために作られる．そうした泡沫の共同体は，不特定多数のユーザが新たに参加することも多い．新たな参加者が参加した結果，多数派や少数派ができる場合がある．少数派が自らの意見を公言せず沈黙してしまい，多数派の意見が支配[14] 的になっていく現象はしばしばある．そうした少数派の参加者，あるいはイノベーター[15] やアーリーアダプター[16] のような参加者は，また，新たなコミュニティや共同体，コミュニケーション方法を作るであろう．コミュニケーションにおける記号の意味作用を分析する理論に，記号理論という分野がある [152]．歴史は古く，古代ギリシャのプラトン[17]やアリストテレスに端を発し，中世のスコラ哲学[18]の伝統を持つ分野である．したがって，記号の意味作用を分析は，古代ギリシャ時代から行われてきたと言える．ゆえに，インターネットのメディアで作られるコミュニティや共同体，コミュニケーション方法は，現時点で有効であっても，泡沫な性質を持つ．一方で，マスメディアや情報配信プラットフォームなど伝播メディアは，公的な制度や情報配信基盤に基づいて成立している．したがって，容易には消滅することはない．伝播メディアを用いたコミュニティケーションは，対応の如何で，大きく結果が変化する．インターネットのメディアは非常にダイナミックな性質を持つが，一時的に大きな影響力があった場合でも，公的な制度や法システムなどと異なり，一過性のブームとなる場合もある．同様に，炎上やフレーミングが発生した場合においても，異なる後続のコミュニティケーションで，一過性のぼやきとして消滅することもある．しかし，情報カスケードのように，価値判断を行う際に，集団全体が画一的な判断になだれ込んでしまう現象もある．炎上やフレーミングが起因し，プライバシーを侵害するコミュニティケーションとなった場合や誹謗中傷に加担した結果となった場合，一個人の問題では済まなくなる．

現状において，感情伝染実験や投票行動に影響力を行使する “デジタルゲリマンダ” なども指摘されている．ソーシャルメディアでターゲットとなった個人は対象になっていることを認識できないまま，誘導され，世論操作などに使用されるような問題なども指摘されている [153]．人々が功利的に利害を追求し，権利を主張し合えば，社会秩序が問題となる．個人に対する誹謗中傷，プライバシーが暴かれる私刑，また，近代を維持するか，放棄するかの転換点にあるという議論もある．画一的な判断になだれ込む性質がある場合，社会的な損失も無視できない規模の問題へ発展する場合もある．提案手法を用いた成果の後続のコミュニケーションが，構造[19]の分析，あるいは抑制すべき情報の検知，社会的な損失の回避となる公的な制度や法システムの基となる発展であること期待したい．

---

[14]行動や意志を統制・束縛することを支配と呼ぶ．

[15]アメリカの社会学者のエヴェリット・M・ロジャースが提唱したイノベーター理論の用語．新たに現れた革新的商品やサービスなどをいち早く受容し，支持する人．

[16]イノベーター理論の用語．新たに現れた革新的商品やサービスなどを比較的早い段階で採用・受容する人．

[17]古代ギリシャの哲学者 (前 427-前 347)，ソクラテスの弟子．

[18]アリストテレスの哲学体系に典拠し，普遍論争および理性と信仰の調和を中心課題にした哲学．スコラ哲学の完成者はトマス=アクィナス．普遍論争はトマス=アクィナスが「普遍は個物の中に実在する」として調停した．

[19]全体を形づくっている種々の構成素による各部分の相互関係によるシステムを静態的・分析的に捉えた表現．対して，構成は立体的・動作的に捉えた表現．

四点目が，コミュニケーションの結果としての発展である．コミュニケーションの目的の一つにモチベーションがあるが，その場合，動機を達成することで，真に利益が得られる結果でなくてはならない．メディアはイメージ効果を重要視している背景から，ここでは，動物行動学や生物学を例に示す．

例えば，モチベーションの目的が優良な個体選別であり，与えるべき動機が，優良な筋肉であるとしよう．発信者は，コミュニケーションにおいて，メッセージは優良な筋肉から構成されている個体の類似記号を送信したとする．類似記号である個体はメッセージである．チャネルである眼を経由して，受信者に情報が伝達される．そして，メッセージの受信の結果から，受信者は類似の個体を選択するが，眼だけでは，優良な筋肉か，あるいは筋肉の模造品なのかは，判別ができない．視覚的に良く見えるだけの個体なのかもしれない．筋肉の模造品や視覚的に良く見えるだけの筋肉を作る方が，安上がりな場合，選択される個体が，都合が良い安上がりに作る方を選択する．そうして作成された個体を，類似記号を用いて優良な個体選別のモチベーションを与えられた受信者が選択した場合，実際には優良でない個体を選択をしたことになる．そうした場合，受信者は個体選択の動機を達成した場合においても，動機の基となった優良な筋肉という利益は得られない．したがって，モチベーションを与えるメッセージが，模造品を明らかにし，優良な筋肉を選択できるメッセージである必要がある．あるいは，模造が困難な優良な筋肉よりも食物獲得能力が高い，異なるモチベーションのメッセージを用いる方が良い場合もある．まつ毛や化粧などがメッセージである場合も同様である．個体がメッセージの場合，受信者は，優良でない個体を判別することで目的が達成できる．しかし，アフォーダンス[20] で明らかであるが，情報の意味解釈は種ごとに異なる．また，功利主義[21]，結果主義[22]，個人主義[23]，無関係対象からの独立性[24] など，受信者は多様な特性を持つ [154]．加えて,「スジを通した」決定でなければ受け入れられない場合もある．ゆえに，受信者や情報路の重要度は，メッセージの作成と比較して，相対的に低い．

雑誌を基に具体例を示す．雑誌におけるメッセージでは，画像やテキスト情報を用いてメッセージを代理する対象を，受信者である解釈者に想起させることを目的とする場合がある．しかし，雑誌は印刷された文字，画像は，対象の類似でしかなく，実際にはインクによって作成された類似の形状や色彩を持つメッセージある．情報伝達の順序や複数の伝播メディアを用いていた場合においても，雑誌によるメッセージをインクとして認識している場合，想起されず，情報路が眼の場合は，モチベーションに効果がない．受信者に対してモチベーションを与えたい場合は，動機や意欲となる要因に相当する刺激が伝達する情報路を明らかにし，効果的な情報路を使用した上で，優れたメッセージを作成する必要がある．優れたメッセージは，受信者のモチベーションやその結果に大きな影響を与える．ゆえに，今後においても，メッセージは重要であると言える．提案手法を用いた成果が，動機や意欲に効果的な受信者に対する有益なメッセージ作成に発展していくと期待したい．

---

[20]環世界が生命体に応じて返す反応．環境の意味や価値は認識主体によって加工されるのではなく，環境からの刺激のうちにすでに提供され，固有の形をとっているという思想．固有は，そのものだけが持っているさま．

[21]自らの効用を最大とする行為を選択すること．功利，効用を生活の究極基準とする主義．

[22]選択肢のもたらす結果に対して決定され，プロセスに影響を受けないとすること．

[23]個人の選好は他の人の選好によって影響されないこと．個人の権利や価値を重んじる主義．

[24]選択肢の部分集合内での選好順序は，集合に含まれない選択肢の増減や選好順序の変更に影響されないこと．

参考文献

[1] 椎塚久雄. 感性工学ハンドブック. 朝倉書店, 2013.

[2] 亀井昭宏. 電通広告事典. 電通, 2008.

[3] 原田健一. 地域の映像とは何か:ローカル局のドキュメンタリー映像の文化的，社会的文脈とその問題. マス・コミュニケーション研究, Vol. 92, pp. 3–21, 2018.

[4] 電通総研. 情報メディア白書 2017. ダイヤモンド社, 2017.

[5] 上村忠. マス・マーケットは崩壊したか? : 分衆・小衆論とテレビ離れ論への批判を中心に (特集 大衆社会論とジャーナリズム). 新聞学評論, Vol. 35, pp. 156–165,286–28, 1986.

[6] 中尾瑞紀. 「愛国心」についての一考察 : 共同体意識をベースにして. 法政論叢, Vol. 38, No. 1, pp. 126–137, 2001.

[7] 荒井紀一郎. 民意のベースライン : 新聞報道による議題設定効果の測定 (特集 民意). 年報政治学, Vol. 2014, No. 1, pp. 104–122, 2014.

[8] 宇佐美誠. 若者に公正な社会. 動向, Vol. 20, No. 4, pp. 54–57, 2015.

[9] 安宅川佳之. 世代間対立の時代の公共政策. 日本福祉大学経済論集, No. 35, pp. 1–30, aug 2007.

[10] 田中東子. オンライン空間と女性たちによる表現文化の分析可能性 (特集女性による表現文化の現在とメディア). マス・コミュニケーション研究, Vol. 83, pp. 75–93, 2013.

[11] 清水裕士, 小杉考司. 対人行動の適切性判断と社会規範― 「社会関係の論理学」の構築―. 実社心研, Vol. 49, No. 2, pp. 132–148, 2010.

[12] 黒岩高明. コミュニケーションにおけるプレゼンテーションとビジュアル化 (講座ビジュアル化プレゼンテーション講座 (1)). 情報の科学と技術, Vol. 36, No. 4, pp. 183–187, 1986.

[13] 乾孝司, 奥村学. テキストを対象とした評価情報の分析に関する研究動向. 自然言語処理, Vol. 13, No. 3, pp. 201–241, jul 2006.

[14] 関洋平. 意見分析コーパスの現状と課題. 情報処理学会論文誌データベース（TOD）, Vol. 6, No. 4, pp. 85–103, sep 2013.

[15] 岩田具治. 潜在トピックモデルを用いたデータマイニング (特集 データを読み解く技術: ビッグデータ,e-サイエンス, 潜在的ダイナミクス) – (潜在的ダイナミクス : 深い変化を読み解く). 電子情報通信学会誌, Vol. 97, No. 5, pp. 405–409, may 2014.

[16] 濱岡豊. バズ・マーケティングの展開. *Ad Studies*, Vol. 20, pp. 5–10, 2007.

[17] 佐藤尚之, 金田育子, 京井良彦, 信澤宏至, 茂呂譲治, 橋口幸生, 宮林隆吉. Sips ～来るべきソーシャルメディア時代の新しい生活者消費行動モデル概念～. http://www.dentsu.co.jp/sips/index.html. 最終閲覧日:2017-08-03.

[18] Mike J. Walker. Hype cycle for emerging technologies, 2017. *Stamford, USA: Gartner*, 2017.

[19] Irem Arikan, Srikanta Bedathur, and Klaus Berberich. Time will tell: Leveraging temporal expressions in IR. *Second ACM International Conference on Web Search and Data Mining (WSDM '09) - Late Breaking Results*, 01.

[20] Steve Chien and Nicole Immorlica. Semantic similarity between search engine queries using temporal correlation. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pp. 2–11, New York, NY, USA, 2005. ACM.

[21] Qiankun Zhao, Steven C. H. Hoi, Tie-Yan Liu, Sourav S. Bhowmick, Michael R. Lyu, and Wei-Ying Ma. Time-dependent semantic similarity measure of queries using historical click-through data. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pp. 543–552, New York, NY, USA, 2006. ACM.

[22] 岩田具治. トピックモデル (機械学習プロフェッショナルシリーズ). 講談社, 2015.

[23] Kleinberg Jon. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, Vol. 7, No. 4, pp. 373–397, 2003.

[24] Jon D. Mcauliffe and David M. Blei. Supervised topic models. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pp. 121–128. Curran Associates, Inc., 2008.

[25] Masayuki Okamoto and Masaaki Kikuchi. Discovering volatile events in your neighborhood: Local-area topic extraction from blog entries. In Gary Geunbae Lee, Dawei Song, Chin-Yew Lin, Akiko Aizawa, Kazuko Kuriyama, Masaharu Yoshioka, and Tetsuya Sakai, editors, *Information Retrieval Technology*, pp. 181–192, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[26] 吉田光男, 荒瀬由紀. トレンドキーワードに関するウェブリソースの横断的分析. 情報処理学会論文誌データベース（TOD）, Vol. 9, No. 1, pp. 20–30, mar 2016.

[27] 中島伸介, 張建偉, 稲垣陽一, 中本レン. 大規模なブログ記事時系列分析に基づく流行語候補の早期発見手法. 情報処理学会論文誌データベース（TOD）, Vol. 6, No. 1, pp. 1–15, jan 2013.

[28] 古川忠延, 松尾豊, 大向一輝, 内山幸樹, 石塚満. ブログ上での話題伝播に注目した重要語判別. 知能と情報 : 日本知能情報ファジィ学会誌 : journal of Japan Society for Fuzzy Theory and Intelligent Informatics, Vol. 21, No. 4, pp. 557–566, aug 2009.

[29] 山家雄介, 中村聡史, アダムヤトフト, 田中克己. ソーシャルブックマーキングの周期性発見と時期連動型検索ランキングへの適用. 情報処理学会論文誌データベース（TOD）, Vol. 2, No. 3, pp. 130–140, sep 2009.

[30] 山家雄介, 中村聡史, アダムヤトフト, 田中克己. ソーシャルブックマークの特性分析とそれに基づく web 検索結果の再ランキング手法. 情報処理学会論文誌データベース（TOD）, Vol. 1, No. 1, pp. 88–100, jun 2008.

[31] 川本貴史, 豊田正史, 吉永直樹. マイクロブログからの社会的影響力を持つ情報カスケードの検知手法. 情報処理学会論文誌データベース（TOD）, Vol. 9, No. 2, pp. 23–33, jun 2016.

[32] 吉川友也, 岩田具治, 澤田宏. ユーザの潜在特徴を考慮したソーシャルネットワーク上の情報拡散モデル. 情報処理学会論文誌データベース（TOD）, Vol. 6, No. 5, pp. 85–94, dec 2013.

[33] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: Quantifying influence on twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pp. 65–74, New York, NY, USA, 2011. ACM.

[34] Justin Cheng, Lada Adamic, P. Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. Can cascades be predicted? In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pp. 925–936, New York, NY, USA, 2014. ACM.

[35] Wojciech Galuba, Karl Aberer, Dipanjan Chakraborty, Zoran Despotovic, and Wolfgang Kellerer. Outtweeting the twitterers-predicting information cascades in microblogs. *WOSN*, Vol. 10, pp. 3–11, 2010.

[36] Hongyu Gao, Yan Chen, Kathy Lee, Diana Palsetia, and Alok N. Choudhary. Towards online spam filtering in social networks. In *NDSS*, 2012.

[37] Qiaozhu Mei, Chao Liu, Hang Su, and ChengXiang Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pp. 533–542, New York, NY, USA, 2006. ACM.

[38] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, Vol. 1, No. 1, May 2007.

[39] 松村真宏, 三浦麻子, 柴内康文, 大澤幸生, 石塚満. 2 ちゃんねるが盛り上がるダイナミズム. 情報処理学会論文誌, Vol. 45, No. 3, pp. 1053–1061, mar 2004.

[40] 鳥海不二夫, 山本仁志, 諏訪博彦, 岡田勇, 和泉潔, 橋本康弘. 大量ＳＮＳサイトの比較分析. 人工知能学会論文誌, Vol. 25, No. 1, pp. 78–89, 2010.

[41] 難波英嗣. 動向情報の抽出と要約ー動向をまとめるー. 日本知能情報ファジィ学会誌, Vol. 22, No. 5, pp. 549–555, 2010.

[42] Zamanian Mostafa and Heydari Pooneh. Readability of texts: State of the art. *Theory & Practice in Language Studies*, Vol. 2, No. 1, 2012.

[43] 藤原郁郎. レクサイル指標の位置づけと計測方法. 関西大学外国語教育フォーラム, No. 15, pp. 41–55, mar 2016.

[44] 小山由紀江. Readability 指標による科学テキストの予備分析. *New directions*, No. 30, pp. 19–35, 2012.

[45] Kincaid J Peter, Fishburne Jr Robert P, Rogers Richard L, and Chissom Brad S. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. 1975.

[46] 柴崎秀子. リーダビリティー研究と「やさしい日本語」. 日本語教育, Vol. 158, pp. 49–65, 2014.

[47] 相澤一美. 読解における語彙カバー率と理解度の関係. 教材学研究, Vol. 22, pp. 23–30, 2011.

[48] 清川英男. アメリカ英語の統計的研究:cbs news の word count. 時事英語学研究, Vol. 1982, No. 21, pp. 19–25, 1982.

[49] Hading Muhaimin, Matsumoto Yuji, and Sakamoto Maki. Japanese lexical simplification for non-native speakers. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, pp. 92–96, 2016.

[50] 芋野美紗子, 吉村枝里子, 土屋誠司. 新聞記事中の難解語を平易な表現へ変換する手法の提案. 自然言語処理, Vol. 20, No. 2, pp. 105–132, jun 2013.

[51] Kajiwara Tomoyuki and Yamamoto Kazuhide. Evaluation dataset and system for japanese lexical simplification. In *Proceedings of the ACL-IJCNLP 2015 Student Research Workshop*, pp. 35–40, 2015.

[52] Kajiwara Tomoyuki, Matsumoto Hiroshi, and Yamamoto Kazuhide. Selecting proper lexical paraphrase for children. In *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*, pp. 59–73, 2013.

[53] 田中英輝, 熊野正, 後藤功雄, 美野秀弥. やさしい日本語ニュースの制作支援システム. 自然言語処理, Vol. 25, No. 1, pp. 81–117, 2018.

[54] Kajiwara Tomoyuki and Komachi Mamoru. Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1147–1158, 2016.

[55] Suzuki Yui, Kajiwara Tomoyuki, and Komachi Mamoru. Building a non-trivial paraphrase corpus using multiple machine translation systems. In *Proceedings of ACL 2017, Student Research Workshop*, pp. 36–42, 2017.

[56] 大谷卓史, Takushi Otani. プライバシーの多義性と文脈依存性をいかに取り扱うべきか : nissenbaum の文脈的完全性と solove のプラグマティズム的アプローチの検討. 吉備国際大学研究紀要（人文・社会科学系）, No. 26, pp. 41–62, mar 2016.

[57] 堀部政男. プライバシー・個人情報保護の新課題. 商事法務, 2010.

[58] 佐藤匡. 住基ネットとプライバシー : マイナンバーにむけて. 地域学論集 : 鳥取大学地域学部紀要, Vol. 12, No. 1, pp. 59–77, aug 2015.

[59] 著作権委員会. 肖像権・撮る側の問題点-二つの顔・プライバシー権とパブリシティ権-. 日本写真家協会会報, Vol. 127, pp. 30–31, 2004.

[60] 佐久間淳. データ解析におけるプライバシー保護 (機械学習プロフェッショナルシリーズ). 講談社, 2016.

[61] 佐久間淳. プライバシー保護データマイニング (私のブックマーク). 人工知能学会誌, Vol. 26, No. 5, pp. 533–536, sep 2011.

[62] 竹井潔. ビッグデータ時代におけるプライバシー : カナダを中心として. 聖学院大学論叢, Vol. 28, No. 1, pp. 33–52, 2015.

[63] 佐生明陽, 輪島幸治, 雨車和憲, 田中勇帆, 嶋田茂, 小河誠巳, 古川利博. Sns 記事の語彙的結束性の分析によるプライバシー関連語の抽出精度の向上. DEIM2015 第 7 回データ工学と情報マネジメントに関するフォーラム講演論文集, E2-1. 電子情報通信学会 データ工学研究専門委員会 他共催, 2015.

[64] 高崎晴夫. パーソナライズド・サービスに対する消費者選好に関する研究:プライバシー懸念の多様性に着目した実証分析. 情報通信学会誌, Vol. 34, No. 3, pp. 25–39, 2016.

[65] 町田史門, 梶山朋子, 嶋田茂, 越前功. Sns におけるセンシティブデータの漏洩検知に基づく公開範囲の設定方式. 情報処理学会論文誌, Vol. 55, No. 9, pp. 2092–2103, sep 2014.

[66] 高田さとみ, 周子胤, 高田美樹, 大本茂史, 岸本拓也, 奈良育英, 嶋田茂. キャプションテキスト感情の分析によるプライバシー侵害シーン抽出. 研究報告自然言語処理（NL）, Vol. 2014, No. 6, pp. 1–5, jan 2014.

[67] 高田さとみ, 小山貴之, 町田史門, 宋洋, 嶋田茂. Sns 画像投稿時に発生するプライバシー侵害の要因分析. 電子情報通信学会技術研究報告. IE, 画像工学, Vol. 112, No. 189, pp. 69–74, aug 2012.

[68] 大本茂史, 岸本拓也, 高田美樹, 高田さとみ, 奈良育英, 周子胤, 嶋田茂. ウェアラブルカメラによる sns 記事投稿時に発生するプライバシー侵害の特徴分析. DEIM2014 第 6 回データ工学と情報マネジメントに関するフォーラム講演論文集, B2-1. 電子情報通信学会 データ工学研究専門委員会 他共催, 2014.

[69] 山口真一. 実証分析による炎上の実態と炎上加担者属性の検証. 情報通信学会誌, Vol. 33, No. 2, pp. 53–65, 2015.

[70] 平井智尚. なぜウェブで炎上が発生するのか : 日本のウェブ文化を手がかりとして. 情報通信学会誌, Vol. 29, No. 4, pp. 61–71, mar 2012.

[71] 藤代裕之. ディジタルゲリマンダとは何か-選挙区割策略からフェイクニュースまで- : 3. ソーシャルメディアと想像の共同体. 情報処理, Vol. 58, No. 12, pp. 1080–1084, nov 2017.

[72] 家辺勝文. インターネットの技術基盤の国際化と日本語文書処理 (特集デジタル時代の日本語). 情報の科学と技術, Vol. 64, No. 11, pp. 450–455, 2014.

[73] 福田誠, 吉田武尚, 吉田誠, 樫岡健史. 中学校技術・家庭科教科書電気領域の表記・表現について. 日本教科教育学会誌, Vol. 23, No. 3, pp. 37–42, 2000.

[74] 西川勇佑, 中村雅子. Line コミュニケーションの特性の分析. 東京都市大学横浜キャンパス情報メディアジャーナル, No. 16, pp. 49–59, apr 2015.

[75] 金明哲. 文節パターンに基づいた文章の書き手の識別. 行動計量学, Vol. 40, No. 1, pp. 17–28, mar 2013.

[76] 石田栄美, 安形輝, 野末道子. 文体からみた学術的文献の特徴分析. 三田図書館・情報学会研究大会発表論文集, pp. 33–36. 三田図書館・情報学会, 2004.

[77] Deerwester Scott, Dumais Susan T, Furnas George W, Landauer Thomas K, and Harshman Richard. Indexing by latent semantic analysis. *Journal of the American society for information science*, Vol. 41, No. 6, pp. 391–407, 1990.

[78] Blei David M., Ng Andrew Y., and Jordan Michael I. Latent dirichlet allocation. *J. Mach. Learn. Res.*, Vol. 3, pp. 993–1022, March 2003.

[79] 高村大也, 松本裕治. Svm を用いた文書分類と構成的帰納学習法. 情報処理学会論文誌データベース（TOD）, Vol. 44, No. 3, pp. 1–10, mar 2003.

[80] 奥村学. マイクロブログマイニングの現在. 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, Vol. 111, No. 427, pp. 19–24, jan 2012.

[81] 大里彩乃. 畳語の研究. 言語文化研究, No. 22, pp. 1–16, mar 2014.

[82] 高橋徹. 社会システム分化とゼマンティク：ルーマンにおける社会変動論の一視角. 社会学評論, Vol. 49, No. 4, pp. 620–634, mar 1999.

[83] 谷口永里子, 高橋真理子. 最新の女性ファッション雑誌における日本語の特徴の量的分析-年代差に焦点をあてて- . 言語処理学会 第 22 回年次大会講演論文集, D5-1. 一般社団法人言語処理学会, 2016.

[84] 佐藤政光. 日本語学習者の語彙習得に関する調査研究 (1) 基本語彙の問題点について. 明治大学人文科学研究所紀要, No. 44, pp. 169–180, feb 1999.

[85] 金庭久美子. 専門用語指導のための選定の試み:―ニュース語彙を例として―. 日本語教育方法研究会誌, Vol. 17, No. 1, pp. 2–3, 2010.

[86] 国立国語研究所. 日本語教育のための基本語彙調査, 1984.

[87] 河合敦夫. 意味属性の学習結果にもとづく文書自動分類方式. 情報処理学会論文誌, Vol. 33, No. 9, pp. 1114–1122, sep 1992.

[88] 松吉俊, 江口萌, 佐尾ちとせ, 村上浩司, 乾健太郎, 松本裕治. テキスト情報分析のための判断情報アノテーション. 電子情報通信学会論文誌. D, 情報・システム, Vol. 93, No. 6, pp. 705–713, jun 2010.

[89] 上岡裕大, 成田和弥, 水野淳太, 乾健太郎. 述部機能表現に対する意味ラベル付与. 研究報告音声言語情報処理（SLP）, Vol. 2014, No. 9, pp. 1–9, may 2014.

[90] 福田一雄. 日本語モダリティ覚え書き (その 1). 言語の普遍性と個別性, No. 5, pp. 1–13, mar 2014.

[91] 西原陽子, 松村真宏, 谷内田正彦. Q&a コミュニティでの質疑応答パターンの理解. 人工知能学会全国大会論文集, pp. 1–4. 人工知能学会, 2008.

[92] 横山友也, 宝珍輝尚, 野宮浩揮, 佐藤哲司. 文章の特徴量を用いた質問回答文の印象の因子得点の推定. 日本感性工学会論文誌, Vol. 12, No. 1, pp. 15–24, 2013.

[93] 細貝亮. メディアが内閣支持に与える影響力とその時間的変化: 新聞社説の内容分析を媒介にして. マス・コミュニケーション研究, Vol. 77, pp. 225–242, 2010.

[94] 中尾桂子. 品詞構成率に基づくテキスト分析の可能性 : メール自己紹介文、小説、作文、名大コーパスの比較から. 大妻女子大学紀要. 文系, Vol. 42, pp. 128–101, mar 2010.

[95] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In *In Proc. of EMNLP*, pp. 230–237, 2004.

[96] 関根聡, 竹内康介. 最新の女性ファッション雑誌における日本語の特徴の量的分析-年代差に焦点をあてて- . 言語処理学会 第 13 回年次大会ワークショップ「言語的オントロジーの構築・連携・利用」, pp. 23–26. 一般社団法人 言語処理学会, 2007.

[97] 藤井裕也, 飯田龍, 徳永健伸. Wikipedia 記事を利用した曖昧性のある表現の固有表現クラス分類. 言語処理学会 第 16 回年次大会講演論文集, pp. 15–18. ”一般社団法人言語処理学会, 2010.

[98] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一. 意見抽出のための評価表現の収集. 自然言語処理, Vol. 12, No. 3, pp. 203–222, 2005.

[99] 東山昌彦. 述語の選択選好性に着目した名詞評価極性の獲得. 言語処理学会 第 14 回年次大会論文集, pp. 584–587. 一般社団法人言語処理学会, 2008.

[100] Takamura Hiroya, Inui Takashi, and Okumura Manabu. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pp. 133–140, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[101] 小林のぞみ, 乾健太郎, 松本裕治. 意見情報の抽出／構造化のタスク仕様に関する考察. 情報処理学会研究報告自然言語処理（NL）, Vol. 2006, No. 1, pp. 111–118, jan 2006.

[102] 岡崎直観. Web 文書からの人の安全・危険に関わる情報の抽出. 言語処理学会 第 18 回年次大会発表論文集, pp. 895–898. 一般社団法人言語処理学会, 2012.

[103] 小谷龍ノ介. 新聞記事における偏向性の定量評価. 法政大学大学院紀要. 理工学・工学研究科編, Vol. 58. 法政大学大学院理工学・工学研究科, mar 2017.

[104] 徳久良子, 乾健太郎, 松本裕治. Web から獲得した感情生起要因コーパスに基づく感情推定. 情報処理学会論文誌, Vol. 50, No. 4, pp. 1365–1374, apr 2009.

[105] 丹治広樹. トラブルを表す文の web からの抽出. 言語処理学会 第 15 回年次大会発表論文集, pp. 140–143. 一般社団法人言語処理学会, 2009.

[106] 野宮浩揮, 宝珍輝尚. 顔特徴量の有用性推定に基づく特徴抽出による表情認識. 知能と情報, Vol. 23, No. 2, pp. 170–185, 2011.

[107] 小林重信, 吉田幸司, 山村雅幸. Ga によるパレート最適な決定木集合の生成. 人工知能学会誌, Vol. 11, No. 5, pp. 778–785, 1996.

[108] 馬場真哉, 松石隆. ランダムフォレストを用いたサンマ来遊量の予測. 日本水産学会誌, Vol. 81, No. 1, pp. 2–9, 2015.

[109] 平博順, 春野雅彦. Support vector machine によるテキスト分類における属性選択. 情報処理学会論文誌, Vol. 41, No. 4, pp. 1113–1123, 2000.

[110] 永田賢二, 岡田真人. スパースモデリングを用いた特徴選択と地球科学データ解析 (特集スパースモデリング: 情報処理の新しい流れ). 応用数理, Vol. 25, No. 1, pp. 5–9, 2015.

[111] 酒井英昭. 主成分分析と独立成分分析 (ソフトコンピューティング特集号). システム／制御／情報, Vol. 43, No. 4, pp. 188–195, 1999.

[112] 亀岡弘和. 非負値行列因子分解. 計測と制御, Vol. 51, No. 9, pp. 835–844, sep 2012.

[113] 安川武彦. 非負値行列因子分解を用いたテキストデータ解析. 計算機統計学, Vol. 28, No. 1, pp. 41–55, 2015.

[114] 永田昌明, 平博順. 情報論的学習理論とその応用 : テキスト分類-学習理論の「見本市」-. 情報処理, Vol. 42, No. 1, pp. 32–37, jan 2001.

[115] 新納浩幸. 決定リストを弱学習器としたアダブーストによる日本語単語分割. 自然言語処理, Vol. 8, No. 2, pp. 3–18, apr 2001.

[116] 飯山将晃. 使える!統計検定・機械学習-iv : Random forests を用いたパターン認識. システム／制御／情報, Vol. 59, No. 2, pp. 71–76, 2015.

[117] 山岡啓介. ランダムフォレスト. 映像情報メディア学会誌 : 映像情報メディア, Vol. 66, No. 7, pp. 573–575, jul 2012.

[118] 船橋賢一. 階層型ニュ-ラルネットワ-クの原理的機能 (ニュ-ラルネット 特集). 計測と制御, Vol. 30, No. 4, pp. p280–284, apr 1991.

[119] 甘利俊一. 自然勾配学習法-学習空間の幾何学. 計測と制御, Vol. 40, No. 10, pp. 735–739, oct 2001.

[120] 寺西大, 大松繁, 小坂利寿. ニューラルネットワークによる音響ケプストラムを用いた紙幣の新旧識別. 電気学会論文誌Ｅ（センサ・マイクロマシン部門誌）, Vol. 119, No. 8, pp. 955–961, 1999.

[121] 浦田正夫. k-nearest neighbours 判別を用いたクラスター解析のバリデーション (種々のモデルの統計的解析). 数理解析研究所講究録, Vol. 1603, pp. 111–119, jun 2008.

[122] Smola Alex J. and Schölkopf Bernhard. A tutorial on support vector regression. *Statistics and Computing*, Vol. 14, No. 3, pp. 199–222, August 2004.

[123] 石川葉子, 水上雅博, 吉野幸一郎, Sakti Sakriani, 鈴木優, 中村哲. 感情表現を用いた説得対話システム. 人工知能学会論文誌, Vol. 33, No. 1, pp. 1–9, 2018.

[124] 小林正幸, 小西康夫, 藤田貞雄, 石垣博行. サポートベクタ回帰モデルを用いた超音波モータの位置決め制御. 精密工学会誌論文集, pp. 596–601. 公益社団法人 精密工学会, 2006.

[125] 本村陽一. ベイジアンネットによる確率的推論技術. 計測と制御, Vol. 42, No. 8, pp. 649–654, 2003.

[126] Rehurek Radim and Sojka Petr. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer, 2010.

[127] Halko N., Martinsson P. G., and Tropp J. A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, Vol. 53, No. 2, pp. 217–288, May 2011.

[128] Hoffman Matthew D., Blei David M., and Bach Francis. Online learning for latent dirichlet allocation. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, NIPS'10, pp. 856–864, USA, 2010. Curran Associates Inc.

[129] Bradford Roger B. An empirical study of required dimensionality for large-scale latent semantic indexing applications. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pp. 153–162, New York, NY, USA, 2008. ACM.

[130] 桑原（中島）尚子. 情報定義に内在する静的視座と動的視座. 日本社会情報学会全国大会研究発表論文集, pp. 192–195. 日本社会情報学会, 2007.

[131] 伊藤守. 情報概念について：主知主義的な枠組みから解き放つために (特集社会情報学からの発信). 社会情報学, Vol. 1, No. 1, pp. 3–19, 2012.

[132] Kensuke Mitsuzawa, Maito Tauchi, Mathieu Domoulin, Masanori Nakashima, and Tomoya Mizumoto. Fkc corpus: a japanese corpus from new opinion survey service. *Proc. of the Novel Incentives for Collecting Data and Annotation from People: types, implementation, tasking requirements, workflow and results*, pp. 11–18, 2016.

[133] 光富健一. デューイ十進分類法 (ddc)(特集 分類について考える). 情報の科学と技術, Vol. 39, No. 11, pp. 478–483, 1989.

[134] 菊地悟. 大学生の外来語意識 (1) : イメージ・表記・語種意識の調査から. 岩手大学教育学部附属教育実践研究指導センター研究紀要, Vol. 4, pp. 61–73, 1994.

[135] 山崎誠, 小沼悦. 現代雑誌における語種構成. 言語処理学会 第 10 回年次大会発表論文集, pp. 670–673. 一般社団法人言語処理学会, 2004.

[136] 堀切友紀子. 外来語に関する研究動向 : 使用意識と言語接触の視点から. お茶の水女子大学人文科学研究, Vol. 9, pp. 113–124, mar 2013.

[137] Apple サポートコミュニティ. https://discussionsjapan.apple.com. 最終閲覧日:2014-01-29.

[138] Stack Exchange Data Dump.
https://archive.org/details/stackexchange
(Publication date 2018-09-05).

[139] Aozora Bunko. http://www.aozora.gr.jp
(Archive date 2017-05-31).

[140] jawiki dump progress. https://dumps.wikimedia.org/jawiki/
(Last dumped on 2018-12-01).

[141] Download - RONDHUIT. http://www.rondhuit.com/download.html#ldcc
(Archive date 2012-09).

[142] 飯田健夫. 感覚情報処理の解明とその社会的貢献. 計測と制御, Vol. 41, No. 10, pp. 692–695, oct 2002.

[143] 小勝俊亘. 小形情報端末を創り出したエレクトロニクス (創立 100 周年記念特集エレクトロニクスが創り出したもの, 創り出すもの) – (エレクトロニクスの現在). 電子情報通信学会誌, Vol. 100, No. 9, pp. 896–901, sep 2017.

[144] 黒川不二雄, 大津智. 情報通信端末としての情報機器のエネルギー技術動向 (小特集次世代を切り開く情報通信エネルギー技術). 電子情報通信学会誌, Vol. 101, No. 4, pp. 355–357, apr 2018.

[145] 鈴木正敏. 光通信及びモバイル通信システムの進化と将来の光無線融合 (小特集マイクロ波・ミリ波フォトニクス技術の新展開). 電子情報通信学会誌, Vol. 101, No. 2, pp. 138–145, feb 2018.

[146] 末松憲治. ワイヤレスインターネット社会を支える高周波集積回路技術 (創立 100 周年記念特集 エレクトロニクスが創り出したもの, 創り出すもの) – (エレクトロニクスの現在). 電子情報通信学会誌, Vol. 100, No. 9, pp. 902–906, sep 2017.

[147] 小口正人, 中野美由紀, 石川佳治, 木俵豊. ビッグデータへの取組みと周辺領域との融合 (創立 100 周年記念特集暮らしを豊かにする情報処理技術) – (研究専門委員会の現在までの道のりとこれから). 電子情報通信学会誌, Vol. 100, No. 10, pp. 1059–1059, Oct 2017.

[148] 奥野克巳. 序 (特集 自然と社会の民族誌-動物と人間の連続性). 文化人類学, Vol. 76, No. 4, pp. 391–397, 2012.

[149] 武田比呂男. 環境・文学・精神史:—— 平野仁啓の仕事に学ぶ ——. 日本文学, Vol. 62, No. 5, pp. 10–19, 2013.

[150] 中根千枝. タテ社会の力学. 講談社, 2009.

[151] 佐野昭. ブラインド信号処理技術. 計測と制御, Vol. 43, No. 6, pp. 521–528, jun 2004.

[152] 白石哲郎. ふたつの記号理論と文化の社会学に関する試論 i. 佛大社会学, Vol. 36, pp. 1–14, mar 2012.

[153] 湯淺墾道. ディジタルゲリマンダとは何か-選挙区割策略からフェイクニュースまで- : 1. ディジタルゲリマンダの法規制の可能性. 情報処理, Vol. 58, No. 12, pp. 1070–1074, nov 2017.

[154] 佐伯胖. 意思決定と社会システム (分散と協調 特集). 計測と制御, Vol. 26, No. 1, pp. p39–44, jan 1987.

謝辞

本論文は，伝播メディアにおけるコミュニケーションを，テキスト情報を主眼に，取り組んだ研究成果を，まとめたものです．本論文の作成にあたり，多数の皆様から有益なご教示，ご指導とご支援をいただきました．

まず，はじめに博士後期課程の主指導教員として，筑波大学 図書館情報メディア系 佐藤 哲司 教授に，終始懇切なご指導を頂いた．ここに深く感謝の意を表する．佐藤 哲司 教授は，当初より，博士論文作成の全過程を通して，卓越した見通しをもち，御指導下さり，論文の構成や進め方を御教授下さった．ご援助無しには本論文は成り立たなかったと思われる．

次に，著者にテキスト情報に関するテーマを与えて下さり，それ以来，研究をご指導頂いた東京理科大学 工学部情報工学科 古川 利博 教授に深厚な謝意を示します．古川 利博 教授は，研究を遂行するにあたり，終始親身なご指導を頂いた．研究室活動においても，多大なご支援，ご指導，有益な御教示を頂きました．ご援助無しには研究が成り立たなかったと言える．

博士後期課程では，副指導教員として，筑波大学 図書館情報メディア系 木暮 啓 教授には，大変丁寧なご指導を頂いた．ここに記し，深謝致します．木暮 啓 教授は，研究課題における複眼的な考察を与えてくださり，加えて，資料などもご推薦頂いた．伝統的なメディアにおける影響力の大きさや，現実課題に対する実際の取組み，潮流などは，木暮 啓 教授のご指導無しには本論文に反映できなかったと言える．また，副指導教員として，ご指導頂いた筑波大学図書館情報メディア系 森嶋 厚行 教授には，学術誌における動向などをご指導頂いた．深く感謝致します．本論文の審査にあたっては，ご担当頂いた筑波大学図書館情報メディア系 芳鐘 冬樹 教授，筑波大学 計算科学研究センター 天笠 俊之 教授には貴重なご助言と激励の言葉を頂いた．深く感謝致します．加えて，本論文の審査をご担当頂いた手塚 太郎 准教授には貴重なご助言に加え，有益な御教示を頂いた．深く感謝致します．

本論文に関しては，一部言語資源を利用した．ここに記して，評価対象データを提供頂いた嶋田 茂 教授，言語資源である livedoor ニュースコーパスを作成頂いた株式会社ロンウイットおよび livedoor ニュース提供元の NHN Japan 株式会社に感謝申し上げます．考察では，株式会社 Insight Tech が国立情報学研究所の協力により，研究目的で提供している「不満調査データセット」から，不満カテゴリ辞書データを利用した．データ提供頂いた株式会社 Insight Tech および国立情報学研究所に感謝申し上げます．加えて，特徴量抽出に用いた辞書の作成者の皆様に感謝致します．そして，本論文ならびに連関した研究において，データ整理のご協力，有益なご討論を頂いた佐藤研究室の皆様，古川研究室の皆様，嶋田研究室の皆様，アンケートにご協力頂いた，大学生，大学院生の皆様，サポート業務経験者の皆様に感謝致します．紙幅の関係で，詳細は省きますが，本論文の作成にご協力頂いた皆様に感謝致します．

全研究業績

博士論文に関する業績

査読付き論文誌

- 輪島 幸治，古川 利博，佐藤 哲司，SNS 記事におけるプライバシー侵害に関わる特徴量の推定，情報社会学会，情報社会学会誌，Vol.13，No.1，2018，pp.19-32
- 輪島 幸治，古川 利博，佐藤 哲司，基底選択を用いた話題性の評価，情報知識学会，情報知識学会誌，(採択決定)

査読付き国際会議

- **Koji Wajima**,Kei Kogure,Toshihiro Furukawa,Tetsuji Satoh, ”Evaluation of Japanese Text Characteristics with Simplified Corpora using Base Selection”,The 20th International Conference on Information Integration and Web-based Applications & Services (iiWAS’18), pp. 46-54, Yogyakarta, Indonesia (Nov. 2018).

研究会・全国大会発表

- 輪島 幸治，木暮 啓，古川 利博，佐藤 哲司，可読性に基づいた日本語テキスト情報の特徴量評価，電子情報通信学会 データ工学研究専門委員会 他共催，第 10 回データ工学と情報マネジメントに関するフォーラム DEIM2018，C2-1，清風荘 福井県あわら市 (Mar. 4-6, 2018)
- 輪島 幸治，古川 利博，佐藤 哲司，非負値行列因子分解とサポートベクタ回帰モデルに基づいた共感された質問記事における特徴抽出手法の提案，情報処理学会，情報処理学会 第 81 回全国大会，1C-01，福岡大学 七隈キャンパス (Mar. 14-16 2019)

その他の論文

査読付き論文誌

- **Koji Wajima**,Kei Kogure,Toshihiro Furukawa,Tetsuji Satoh, ”Extract of Japanese Text Characteristics of Simplified Corpora using Non-negative Matrix Factorization”, Journal of Data Intelligence (JDI), Rinton Press, USA, (in press)

査読付き国際会議

- **Koji Wajima**,Tetsuji Satoh, ”Urgent Question Detection based on the Review Points and Sentiment Words”, The 17th International Conference on Information Integration and Web-based Applications & Services (iiWAS'15), pp. 307 - 311, Brussels, Belgium (Dec. 2015)

研究会・全国大会発表

- 輪島 幸治，木暮 啓，佐藤 哲司，年次報告書を用いた時系列トピック分析 : 環境省の環境白書を事例にして，電子情報通信学会技術研究報告，電子情報通信学会，2017-06-23，117，108，31-36

- 輪島 幸治，佐藤 哲司，単語の類似度と感情表現を考慮した質問記事の判定手法，電子情報通信学会 他共催，第 8 回データ工学と情報マネジメントに関するフォーラム DEIM2016，B3-2，ヒルトン福岡シーホーク (Feb. 29-Mar. 2, 2016)

- 輪島 幸治，佐藤 哲司，評価視点と感情表現に基づく質問記事重要度判定手法の提案，マルチメディア，分散，協調とモバイル（DICOMO 2015）シンポジウム

- 隅岡 隆之，尾崎 敏司，村上 陽子， 超穎，畠山 智美，輪島 幸治，嶋田 茂，ウェアラブルカメラ撮影者へのコンテキストベースプライバシー侵害理由の警告サービスシステム，第 7 回データ工学と情報マネジメントに関するフォーラム DEIM2015，G2-2，磐梯熱海ホテル華の湯 (Mar. 2-4, 2015)

- 輪島 幸治，古川 利博，嶋田 茂，KeyGraph による主張点の極性評価〜LDA の潜在トピックを用いて〜，電子情報通信学会 他共催，第 7 回データ工学と情報マネジメントに関するフォーラム DEIM2015，F5-1，磐梯熱海ホテル華の湯 (Mar. 2-4, 2015)

- 佐生 明陽，輪島 幸治，雨車 和憲，田中 勇帆，嶋田 茂，小河 誠巳，古川 利博，SNS 記事の語彙的結束性の分析によるプライバシー関連語の抽出精度の向上，電子情報通信学会 他共催，第 7 回データ工学と情報マネジメントに関するフォーラム DEIM2015，E2-1，磐梯熱海ホテル華の湯 (Mar. 2-4, 2015)

- 堀内 佑城，**輪島 幸治**，古川 利博，潜在的ディリクレ法におけるラベリング自動化，電子情報通信学会 他共催，第 7 回データ工学と情報マネジメントに関するフォーラム DEIM2015，D1-4，磐梯熱海ホテル華の湯 (Mar. 2-4, 2015)

- 佐生 明陽，**輪島 幸治**，小河 誠巳，嶋田 茂，古川 利博，ステミングと N-gram 共起によるプライバシー関連語の抽出精度向上，研究報告知能システム（ICS），一般社団法人情報処理学会，2014-12-08，2014，12，1-6

- 尾崎 敏司，**輪島 幸治**，隅岡 隆之，村上 陽子， 超穎，嶋田 茂，SNS 動画像投稿記事のトピック抽出とそのカテゴリー化によるプライバシー侵害理由の推定方式，第 5 回 WI2 研究会開催報告，ARG WI2 研究会，2014-11-22

- **輪島 幸治**，小河 誠巳，古川 利博，嶋田 茂，共起語の評価極性に着目したネガティブトピックの評価 (データ工学)，電子情報通信学会技術研究報告，一般社団法人電子情報通信学会，2014-06-21，114，101，73-78

- **輪島 幸治**，小河 誠巳，古川 利博，嶋田 茂，潜在的ディリクレ配分法を用いたネガティブ要因分析，電子情報通信学会 他共催，第 6 回データ工学と情報マネジメントに関するフォーラム DEIM2014，A9-3，淡路夢舞台＆ウェスティン淡路 (Mar. 3-5, 2014)

- **輪島 幸治**，小河 誠巳，古川 利博，ヘルプデスクにおける製品事故に関するクレーム情報の分類：感情極性と製造物責任法に関する訴訟情報から (情報セキュリティ)，電子情報通信学会技術研究報告，一般社団法人電子情報通信学会，2013-11-28，113，326，9-14

- **輪島 幸治**，小河 誠巳，古川 利博，テキスト評価分析を用いたヘルプデスク効率化手法の提案，経営情報学会　全国研究発表大会要旨集，一般社団法人経営情報学会，2013，141-144

- **輪島 幸治**，林 正樹，古川 利博，T2V による MMD 向け CG オブジェクトの再利用可能性の考察 (ポスター (CG), 映像表現・芸術科学フォーラム 2013)，映情学技報，一般社団法人 映像情報メディア学会，2013，37，85-88